Universidad Autónoma de Madrid
Departamento de Biología Molecular
Facultad de Ciencias

Estudio de la evolución estructural en familias de proteínas y su aplicación al refinado de modelos obtenidos por homología

Tesis Doctoral

Alejandra Leo Macías

Director de Tesis
Dr. Ángel Ramirez Ortiz

MADRID 2006

El viaje más largo comienza con un solo paso
Lao-Tse, 604 aC – 531 aC

A mis padres, hermanos,
demás familia y amigos

Agradecimientos

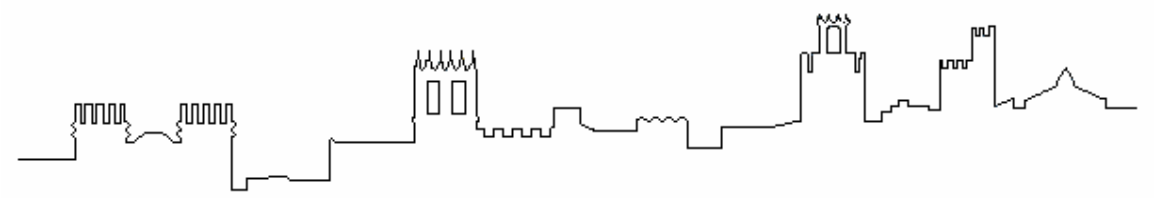
Esta tesis ha sido realizada en el Centro de Biología Molecular Severo Ochoa (CSIC-UAM) con la dirección del Dr. Ángel Ramírez Ortiz, a quien agradezco el haberme introducido en el mundo de la bioinformática, mostrarme la importancia y belleza de las estructuras de proteínas y proporcionarme los medios necesarios para desempeñar el trabajo, que ha sido financiado con una beca de postgrado para la Formación de Personal Investigador (BIO2001-3745), que agradezo también, pues me ha permitido vivir en Madrid durante estos cuatro años dedicándome a lo que me gusta.

Además, esto no habría sido posible sin la colaboración de tantas personas que directa o indirectamente han conseguido que este barco arribara a buen puerto. Desde mis compañeros de laboratorio (actuales y pasados), gracias a David, Ugo, Antonio, Fernando, Rubén G., Rubén admin, Enrique, Virginia, con los que ha sido un auténtico placer compartir el día a día; gracias especiales a Pedro y Han, por sus explicaciones y enseñanzas. A mi amigo Jesús, compañero de tantos metaanálisis y a Edu, por todos los nicaraguas, las risas y los ánimos. A los visitantes, Mary, Dani y Florian, por su alegría. Gracias inmensas a Paulino, por su apoyo y ayuda siempre que lo necesité y por desfacer algunos entuertos. A Reyes y especialmente a Mada, por su colaboración (¡gracias!). De mis visitas a NYC, a András y su grupo, Narcís, Carlos, Edu, Dimtry, por eneñarme otra forma de hacer ciencia, y a mi familia americana, Marta Murcia, Marta Sánchez, Dima, Luis y también a Sergio, por tantos momentos. A las Padilla´s, mis queridas amigas y compañeras de piso, las de ahora, Carmen, Bea, Carolina y María Fernanda, y las de antes, Lili, Rose, Pilar y Gema, que me hicieron sentir tan bien en Madrid. A mis amigos de Badajoz, Silvia, Mara, Conchi, Manolo, Bea, Cristina, Ana, María, Pipo y Meli por su cariño y ánimos a pesar de no entender muy bien qué hacía aquí con tanta “molécula de la vida”.

Gracias también a muchos profesores, investigadores y científicos que en cursos, seminarios o charlas siguieron alimentando la ilusión con sus lecciones de ciencia y de vida.

Por último, mi familia, mis hermanos y mis padres, gracias infinitas a ellos porque sin su amor y apoyo constante nada habría sido posible.

Muchas gracias a todos, el resultado de este trabajo es vuestro.



ÍNDICE

Índice

Índice de contenidos

Índice de algoritmos

Índice de tablas

Índice de figuras

ABREVIATURAS

Abreviaturas

3D	Espacio tridimensional.
ADN	Ácido desoxirribonucleico.
ANM	*Anisotropic Network Model* – Modelo de red anisotrópico.
CAFASP	*Critical Assessment of Fully Automated Structure Prediction Techniques* – Evaluación crítica de técnicas completamente automáticas para la predicción de estructura de proteínas.
CAMPASS	*CAMbridge database of Protein Alignments organised as Structural Superfamilies* – Base de datos de Cambridge de alineamientos estructurales organizadas en superfamilias.
CASP	*Critical Assessment of Techniques for Protein Structure Prediction* – Evaluación crítica de técnicas para la predicción de estructura de proteínas.
CATH	*Class Architecture Topology Homology* – Base de datos para clasificar los dominios proteinicos por clase, arquitectura, topología y homología.
EM-PCA	*Expectation Maximization Principal Component Analysis* – Análisis de componentes principales mediante expectación-maximización.
HMM	*Hidden Markov Model* – Modelo oculto de Markov.
HOMSTRAD	*HOMologous STRucture Alignment Database* – Base de datos de alineamientos estructurales de proteínas homólogas.
LES	*Locally Enhanced Sampling* – Muestreo localmente enriquecido.
LCR	*Local Constraint Refinement* – Refinado usando restricciones locales.
MC	Monte Carlo.
MD	*Molecular Dynamics* – Dinámica molecular.
NMA	*Normal Mode Analysis* – Análisis de modos normales.
PCA	*Principal Component Analysis* – Análisis de componentes principales.
PDB	*Protein Data Bank* – Base de datos que almacena las estructuras a alta resolución de proteínas.
PME	*Particle Mesh Ewald* – Red de partículas de Ewald.
PSSM	*Position Specific Scoring Matrix* – Matriz de puntuación específica de posición.
RCR	*Reduced Contact Refinement* – Refinado usando restricciones de contactos.
RDC	*Residual Dipolar Coupling* – Acoplamiento dipolar residual.

REMC	*Replica Exchange Monte Carlo* – Monte Carlo con intercambio de réplicas.
RMN	Resonancia Magnética Nuclear
RMSD	*Root Mean Squared Distance* – Desviación cuadrática media.
RMSIP	*Root Mean Square Inner Product* – Desviación cuadrática media del producto escalar.
R-X	Cristalografía de rayos X
SAXS	*Small Angle X-ray Scattering* – Dispersión de ángulo pequeño de rayos-X.
SCOP	*Structural Classification of Proteins* – Clasificación estructural de proteínas.
SCS	*Stereochemical Check Score* – Valor de puntuación estereoquímica.
STAMP	*STructural Alignment of Multiple Proteins* – Alineamiento estructural múltiple de proteínas.
UPGMA	*Unweighted Pair Groups Method using Arithmetic Averages* – Método de pares de grupos no ponderado que usa medias aritméticas.
URMS	*Unit-vector Root Mean Square* – Raíz cuadrada del vector unitario.

RESUMEN

Resumen

El refinado estructural de proteínas continúa siendo un reto importante en el campo de la predicción estructural. La mayoría de los intentos de refinar modelos conducen a su degradación, en lugar de a la mejora de su calidad, de manera que muchos protocolos omiten este paso final. Incluso en ausencia de errores en los alineamientos y usando las plantillas óptimas, se ha demostrado que los métodos de modelado basados en patrones tienen limitaciones intrínsecas, lo que sugiere la necesidad de desarrollar otras metodologías si el objetivo es mejorar la calidad final de los modelos propuestos. Las dificultades del refinado estructural se derivan del delicado balance de fuerzas en el estado nativo de las proteínas, que todavía no es reproducible en toda su extensión mediante los campos de fuerza actuales, y de la necesidad de muestrear un gran número de conformaciones alternativas en la búsqueda del mínimo global de energía. En esta tesis se aborda esta segunda cuestión. Se presenta un nuevo algoritmo de alineamiento estructural múltiple, MAMMOTH-mult, que permite detectar las regiones estructuralmente conservadas en familias de proteínas y se estudia su plasticidad mediante análisis de componentes principales y de modos normales. Esto permite caracterizar las deformaciones más importantes que experimentan las estructuras a lo largo de la evolución y las debidas a su propia topología. Se observa que cada familia de proteínas homólogas presenta un patrón de evolución estructural característico, que está fundamentalmente relacionado con la propia topología de la estructura y no con los detalles de la secuencia. Estos patrones de deformación se utilizan para ayudar a facilitar el problema del muestreo en el refinado. Se observa que se puede resolver este problema de manera esencial para la cadena principal de las estructuras definiendo un subespacio pequeño, de unas 50 dimensiones, consistente en una combinación de direcciones favorecidas por la evolución, definidas por los componentes principales de la variación estructural dentro de las familias de proteínas homólogas, y las direcciones de vibración derivadas del análisis de sus modos normales. La mayoría de los centros estructurales de las proteínas en este subespacio combinado se puede representar con menos de 1 Å de RMSD con respecto a sus posiciones correctas. También se muestra que las optimizaciones de intercambio de réplicas de Monte Carlo son muy eficientes para encontrar el mínimo global en este subespacio. Finalmente, se discuten las aplicaciones de esta metodología.

Abstract

Structural refinement of protein models remains as a particularly challenging problem in protein structure prediction. Most attempts to refining comparative models lead to degradation rather than improvement in model quality, so most current comparative modelling procedures omit the refinement step. However, it has been shown that even in absence of alignment errors and using optimal templates, template-only methods have intrinsic limitations, suggesting that other methodologies must be developed if accuracy is ultimately to be improved. It is thought that these difficulties originate from the delicate balance of forces in the native state and the requirement to sample a large number of alternative tightly packed conformations in the search for the global minimum. Here we address this second issue. We present a new algorithm, MAMMOTH-mult, for multiple structural alignment, that allows to detect structural conserved regions in protein families. Applying principal components and normal mode analysis to these regions allows the caracterization of the most important deformations that structures experiment along the evolution and those which are due to their own topologies. We find that each family of homologous proteins has a characteristic template of structural evolution related to its own structure topology rather than to sequence details. We use this information for helping to solve the sampling problem. We show this problem can be essentially solved at the backbone level by defining a small sampling subspace, of 50 dimensions at most, consisting on a combination of evolutionarily favoured directions defined by the principal components of structural variation within a family of homologous proteins and their topological vibrational directions derived from normal mode analyses. Most protein cores in this combined space can be represented within 1 Å accuracy. We also show that Replica Exchange Monte Carlo optimizations in this subspace are very efficient at finding the global minimum neighbourhood in realistic conditions of roughness of the energy landscape. Applications of this methodology are finally discussed.

INTRODUCCIÓN

1. Introducción

El descubrimiento de la estructura molecular del ácido desoxirribonucleico (ADN), por Watson, Crick y Franklin en 1953 (Watson and Crick, 1953) definió el paradigma de la bioquímica y la biología molecular modernas, y estableció la gran importancia de la estructura molecular para el entendimiento de la función celular. Ello estimuló la dedicación de grandes esfuerzos para resolver estructuras de proteínas, los cuales culminaron a finales de la década, cuando Kendrew y Perutz resolvieron las estructuras tridimensionales a alta resolución de la mioglobina (Kendrew et al., 1958) y la hemoglobina (Perutz, 1960). Desde entonces, gracias a los importantes descubrimientos y desarrollos técnicos que se han ido produciendo (como el uso de las enzimas de restricción, la reacción en cadena de la polimerasa, o más recientemente, el uso de la dinámica molecular restringida, la radiación sincrotrón, etc), se han determinado las secuencias y las estructuras de un gran número de proteínas y otras moléculas biológicas.

Con la publicación del genoma humano (Lander et al., 2001), y la generalización de las técnicas de genómica funcional, se abre definitivamente el paso a la era genómica, confirmando la transición desde un concepto reduccionista de la biología hacia una nueva biología integrativa. La idea subyacente es que los sistemas biológicos son más que la suma de sus partes (de las que además se dispone de una gran cantidad de información), por lo que si el objetivo es entender y manipular los mecanismos últimos de la vida, se hace necesario comprender todo el juego de interacciones y relaciones entre sus constituyentes.

En este contexto, se hace necesario el uso de aproximaciones computacionales eficientes que permitan manejar, analizar, modelar e integrar los datos disponibles. Ello ha dado lugar al florecimiento de la Bioinformática, que en su vertiente estructural, es la subdisciplina que nos permite utilizar la información disponible de los seres vivos a escala atómica y subcelular para tratar de entender, modelar y modificar los procesos que tienen lugar en ellos y proporciona el marco en el que se desarrolla el trabajo realizado en esta tesis.

1.1. Importancia de la estructura tridimensional de proteínas

La mayor parte de todo lo relacionado con la estructura y la función de las células tiene que ver con proteínas. Estas moléculas grandes y complejas son polímeros constituidos por secuencias de aminoácidos y la enorme versatilidad que exhiben les permite llevar a cabo muchas funciones esenciales para la vida, ya que, en contraste con la mayoría de los polímeros sintéticos, en los que las moléculas individuales pueden adoptar conformaciones muy diferentes, una proteína usualmente existe en un único estado nativo, que se da bajo condiciones fisiológicas (disoluciones acuosas a pH cercano al neutro y a temperaturas de unos 20-40 ºC). Cualquier consideración acerca de la función de una proteína debe basarse en el entendimiento de su estructura, que a su vez está determinada únicamente por su secuencia de aminoácidos

(Anfinsen, 1973). La estructura tridimensional de las proteínas permite la disposición de determinados grupos químicos en sitios específicos del espacio y es esta precisión la responsable de la gran diversidad de funciones que presentan: las proteínas actúan como catalizadores (enzimas) de una impresionante variedad de reacciones químicas, pero también desempeñan un papel fundamental en funciones estructurales, de transporte, y de regulación en los organismos.

1.2. Determinación estructural

Básicamente, existen dos métodos experimentales que permiten obtener la estructura a alta resolución de una proteína (es decir, el conocimiento de la posición más probable de todos y cada uno de los átomos que la forman). Se trata de la **cristalografía de rayos X** (R-X), y la **resonancia magnética nuclear** (RMN). Además de estos dos métodos de alta resolución, en algunos casos se pueden emplear técnicas complementarias que permiten obtener determinados datos estructurales parciales o de baja resolución, como son el dicroísmo circular, el SAXS o la microscopía electrónica.

Aunque bastante poderosas, estas técnicas presentan una serie de limitaciones que restringen su aplicabilidad y ralentizan la velocidad de acumulación de estructuras experimentales en las bases de datos. Por ejemplo, muchas proteínas son simplemente demasiado grandes como para poder analizarse mediante RMN (actualmente, esta técnica sólo es aplicable a proteínas de unos 30 kDa como máximo, aunque esta cifra va en aumento), o bien son muy difíciles (cuando no, imposibles) de cristalizar. Como consecuencia de estas limitaciones, el número de estructuras de proteínas resueltas experimentalmente de que se dispone en la actualidad resulta pequeño (~ 35.000 estructuras en el PDB (Deshpande et al., 2005)), si se compara con el número de secuencias de proteínas disponibles (~ 2,1 millones de secuencias en UniProt (Bairoch et al., 2005)).

1.3. Bases de datos de estructuras

Para hacer las estructuras accesibles a toda la comunidad científica, éstas se depositan y almacenan inmediatamente después de su publicación en el **PDB**, que es el repositorio que contiene todas las estructuras tridimensionales de proteínas, ácidos nucleicos, carbohidratos y una gran variedad de otros complejos determinados experimentalmente. Los datos primarios que almacena son las coordenadas cartesianas, los grados de ocupación, y los factores de temperatura para todos los átomos de las estructuras, además de otros datos, como referencias a la bibliografía, autores y detalles del experimento, enlaces a otras bases de datos de secuencia, función, etc.

A partir de este repositorio, se han desarrollado diferentes criterios de clasificación de proteínas, cuya aplicación da lugar a distintas bases de datos. Las más importantes y utilizadas son SCOP (Andreeva et al., 2004; Lo Conte et al., 2000; Lo Conte et al., 2002; Murzin et al., 1995), ASTRAL (Brenner et al., 2000; Chandonia et al., 2004; Chandonia et al., 2002) (basada en SCOP) y CATH (Pearl et al., 2005), aunque existen muchas otras (FSSP (Holm and Sander, 1994; Holm and Sander, 1996; Holm and Sander, 1997), 3Dee(Siddiqui et al., 2001), CAMPASS (Sowdhamini et al., 1998)…). En esta tesis se utilizaron **SCOP** y **ASTRAL** por el tratamiento especialmente cuidadoso que presentan de las relaciones evolutivas del universo de proteínas y además de éstas, también se hizo un uso puntual de **HOMSTRAD** (Mizuguchi et al., 1998b) y **CAMPASS** (ver apartado de Métodos, pág. 27).

1.3.1. SCOP

La base de datos de SCOP (*Structural Classification of Proteins*), fue creada y está mantenida por Murzin et al., en Cambridge. El método usado para construirla se basa en la inspección y comparación visual de estructuras.

La unidad de clasificación es el **dominio**. Éste es una región proteínica que tiene su propio centro estructural (*core*) hidrofóbico y relativo aislamiento del resto de la proteína, lo que la hace estructuralmente semi-independiente. Se le considera la unidad de evolución, estructura y función de las proteínas. Normalmente, los dominios son colineales en secuencia, lo que ayuda a identificarlos, pero de manera ocasional, pueden involucrar dos o más regiones de secuencia de una o más cadenas polipeptídicas que no son colineales.

Las proteínas pequeñas usualmente tienen un único dominio y se tratan como tal. Los dominios de las proteínas más grandes se clasifican de forma independiente. SCOP establece una clasificación jerárquica, atendiendo a criterios evolutivos. En el nivel de **familia** se agrupan aquellos dominios que comparten un claro origen evolutivo común, con alta identidad en secuencia y gran similitud de sus estructuras y funciones. Las **superfamilias** a su vez, están compuestas de familias, cuyos miembros comparten una estructura y función parecidas aunque no presentan una similitud en secuencia significativa, pero debido a las relaciones funcionales hay razón para creer que esas familias están evolutivamente relacionadas. Los **plegamientos** (*folds)* consisten en una o más superfamilias que tienen un centro estructural común (es decir, los mismos elementos de estructura secundaria en el mismo orden, con las mismas conexiones topológicas), pero no comparten semejanzas funcionales o de secuencia obvias y por tanto, no hay evidencia de relación evolutiva. Finalmente y dependiendo del tipo de organización de los elementos de estructura secundaria, los plegamientos se agrupan en cuatro **clases** principales: todo-α, todo-β, α/β y α+β. Además, hay otras muchas clases de proteínas atípicas y difíciles de clasificar, son las proteínas de membrana y superficie celular, proteínas pequeñas, proteínas de cuerda enrollada, estructuras de baja resolución, péptidos y proteínas de diseño. Las proteínas-

multidominio están formadas por diferentes dominios pertenecientes a distintas clases, pero que sin embargo, siempre aparecen juntos. La clase de proteínas pequeñas tiene estructuras estabilizadas por puentes disulfuro o por ligandos metálicos en lugar de por centros hidrofóbicos. Las proteínas de membrana, con frecuencia presentan estructuras peculiares a consecuencia de su entorno especial y por eso se agrupan en una clase diferente.

La versión actual de SCOP es la 1.69 y corresponde a julio de 2005; está basada en una entrega de PDB con fecha 1 de octubre de 2004. En la Tabla 1 se presentan los datos correspondientes a esta última versión.

Clase	***# plegmt.***	***# superfamilias***	***# familias***
Todo α	218	376	608
Todo β	144	290	560
α /β	136	222	629
α+β	279	409	717
Proteínas-multidominio	46	46	61
Proteínas de membrana y superficie celular	47	88	99
Proteínas pequeñas	75	108	171
Total	945	1539	2845

Tabla 1. Clasificación de SCOP: Datos estadísticos correspondientes a la versión 1.69 de SCOP. El número total de dominios de proteínas es 70859; se excluyen ácidos nucleicos y modelos teóricos. Tomado de http://scop.mrc-lmb.cam.ac.uk/scop/

1.3.2. ASTRAL

El compendio de ASTRAL, proporciona un conjunto de bases de datos y herramientas útiles para analizar estructuras de proteínas y sus secuencias. Está parcialmente derivada (y aumentada) de SCOP. Al igual que ésta, la mayoría de los recursos que proporciona dependen de los ficheros de coordenadas mantenidos y distribuidos por el PDB. La principal ventaja de ASTRAL es que proporciona un conjunto no redundante de proteínas que corresponden a dominios únicos definidos por SCOP. Esta información es muy útil para el análisis de relaciones evolutivas entre dominios. También sirve para reducir la redundancia presente en el PDB, ya que las secuencias se pueden filtrar para diferentes grados de identidad, dejando un representante conforme a la estructura de mayor calidad del grupo de estructuras redundantes. Cuando se hace tal reducción, un aspecto muy importante a tener en cuenta es cómo escoger un representante de cada grupo de estructuras redundantes. Todas las aproximaciones emplean una lista inicial ordenada de estructuras basada en parámetros de calidad ampliamente usados para las estructuras de rayos X (la resolución y el factor-R). ASTRAL usa además el parámetro SCS (*Stereochemical Check Score*), combinando puntuaciones de PROCHECK (Laskowski et al., 1993) y WHATCHECK (Hooft et al., 1996), dos programas bien conocidos para la determinación de la calidad estereoquímica de las estructuras. Basándose en esta puntuación combinada se escogen estructuras representativas de otras a diferentes valores de corte para el

porcentaje de identidad en secuencia. ASTRAL también proporciona acceso a conjuntos no redundantes de proteínas filtrados por clasificación de SCOP, esto es, correspondientes a los niveles de clase, plegamiento, superfamilia y familia.

1.3.3. HOMSTRAD

HOMSTRAD (*HOMologous STRucture Alignment Database*), es una base de datos de alineamientos basados en estructura para familias de proteínas homólogas. Las secuencias de los miembros representativos de cada familia se alinean basándose en sus estructuras tridimensionales usando los programas STAMP (Russell and Barton, 1992) y COMPARER (Sali and Blundell, 1990). Estos alineamientos basados en estructura se anotan con JOY (Mizuguchi et al., 1998a) y se examinan individualmente.

1.3.4. CAMPASS

CAMPASS (*CAMbridge database of Protein Alignments organised as Structural Superfamilies*), es una compilación de alineamientos de secuencias basados en estructura para dominios de proteínas clasificados en superfamilias según SCOP. En la mayoría de los casos se emplea un valor de corte de 95% de identidad en secuencia para eliminar redundancias. El alineamiento de los miembros de la superfamilia se basa en la conservación de características estructurales, como accesibilidad al solvente, enlaces de hidrógeno y estructura secundaria usando COMPARER (Sali and Blundell, 1990)

1.4. Espacio de estructuras. Evolución estructural

La clasificación de la población del universo de proteínas en un contexto taxonómico jerárquico como se ha descrito en el apartado anterior, ha permitido la obtención de información muy útil a la hora de entender su evolución. Así, se ha visto que a pesar del número prácticamente ilimitado de posibles secuencias, el número de plegamientos básicos que las proteínas presentan en la naturaleza no sólo es finito, sino que además es relativamente pequeño, se estima que probablemente no sean más de 10^3-10^4 (Andreeva et al., 2004; Pearl et al., 2005). Además, la distribución de proteínas en estos plegamientos es altamente heterogénea, de modo que la mayoría de ellas se concentran en unos pocos posibles, mientras que otros plegamientos se dan raramente. Se ha observado que la distribución de las familias codificadas en la lista de plegamientos conocidos en diversos genomas presentan propiedades matemáticas similares y siguen leyes exponenciales asintóticas asociadas con las llamadas redes libres de escala (*scale-free networks*); es decir, redes en las que el número de nodos a los que está conectado un nodo dado, sigue una ley exponencial. Las distribuciones son de la forma $f(i) \sim i^{-\gamma}$, donde $f(i)$ es la frecuencia de los plegamientos que incluyen exactamente i familias y γ es un

parámetro que típicamente asume valores entre 1 y 3 (Zhang and DeLisi, 1998). Este tipo de distribuciones aparecen en otros muchos contextos biológicos (y no biológicos también), como por ejemplo en el número de conexiones entre dominios y proteínas multidominio, etc. La interpretación es que la mayoría de los genes de un genoma únicamente codifican unos pocos tipos de plegamiento, lo que sugiere que la evolución genómica se produce a través de mecanismos extremadamente generales basados en el principio de "proliferación preferencial" (Koonin et al., 2002). Metafóricamente, esto puede traducirse en una situación en donde "el rico se hace más rico". Estas observaciones son la base de las iniciativas de genómica estructural que suponen un esfuerzo internacional para determinar la estructura tridimensional de un representante de cada familia de proteínas (Chandonia and Brenner, 2006).

Por otra parte, el comportamiento observado hasta el momento resulta muy útil en el caso concreto de la predicción estructural en proteínas, ya que ha permitido el desarrollo de diferentes estrategias de modelado explotando el concepto de "*plantilla*". En general, estas estrategias se apoyan en dos hechos fundamentales concretos:

1. A lo largo de la evolución, la estructura de las proteínas es más estable y sus cambios más lentos que los de la secuencia correspondiente, de manera que secuencias similares adoptan estructuras prácticamente idénticas y secuencias relacionadas remotamente, todavía presentan estructuras similares (Lesk and Chothia, 1980).
2. Los dominios en las secuencias proteínicas pueden agruparse en un número relativamente pequeño de familias de dominios con secuencias y estructuras similares (Chandonia and Brenner, 2005; Vitkup et al., 2001). Por ejemplo, el 75-80% de las secuencias de la base de datos UniProt se agrupan en menos de 15.000 familias de dominios (Bateman et al., 2004; Mulder et al., 2005), y de manera similar, todas las estructuras en el PDB se han clasificado en unos $10^3 – 10^4$ plegamientos diferentes (Andreeva et al., 2004; Pearl et al., 2005).

1.5. Predicción de estructura

Dada la gran diferencia entre el número de estructuras y secuencias disponibles en las bases de datos, la predicción o modelado estructural resulta vital para superar el enorme desfase existente. El objetivo final del modelado de proteínas es poder predecir su estructura a partir de su secuencia, con una precisión comparable a los resultados alcanzados experimentalmente. Esto permitiría usar de manera rápida y segura, los modelos de proteínas generados *in sílico* en todos los contextos donde sólo las estructuras experimentales proporcionan una base sólida hoy en día: diseño de fármacos basado en estructura, análisis de la función y de las interacciones, y diseño racional de proteínas con mayor estabilidad o con funciones nuevas potencialmente interesantes.

1.5.1. Estrategias para la predicción estructural

Existen diferentes estrategias de predicción de estructura de proteínas, que se pueden dividir de manera general en dos grandes categorías: métodos basados en el uso de plantillas y métodos que no las usan. Al primer grupo pertenecen el modelado por homología (*homology* o *comparative modeling*) (Marti-Renom et al., 2000) y el reconocimiento de plegamiento (*fold recognition* o *threading*) (Godzik, 2003), y al segundo, los métodos *ab initio* o *de novo* (Bonneau and Baker, 2001). El uso de un método u otro, depende, a grandes rasgos, del porcentaje de identidad en secuencia entre la proteína problema (diana o *target*), cuya estructura se quiere determinar, y la proteína patrón (plantilla o *template*), cuya estructura es conocida. Los límites entre las categorías son borrosos y muchas veces se producen solapamientos entre ellas, pero de manera muy general, se puede considerar que cuando existe un grado de semejanza significativa entre las secuencias de la diana y de las posibles plantillas de la base de datos de estructuras, el modelado por homología/reconocimiento de plegamiento es la técnica de elección, y los métodos *ab initio* en caso contrario.

1.5.2. Modelado por homología

A pesar del reciente progreso que han experimentado los métodos *ab initio* (Bradley et al., 2005) (especialmente para proteínas pequeñas con menos de 100 aminoácidos), el modelado por homología, cuando resulta aplicable, continúa siendo el método más fiable para predecir la estructura tridimensional de una proteína con una calidad comparable a las estructuras de baja resolución determinadas experimentalmente (Baker and Sali, 2001). Esta técnica se ha beneficiado mucho de la continua progresión de las iniciativas de genómica estructural (Sanchez et al., 2000). Como ya se ha comentado, el objetivo de éstas es conseguir una representación estructural significativa del espacio de secuencias, de manera que la mayoría de las restantes se encuentren dentro de una "*distancia de modelado*" razonable (p.ej., > 30% de identidad en secuencia), con respecto a una proteína de estructura conocida (Chandonia and Brenner, 2005; Sanchez et al., 2000; Vitkup et al., 2001). Se estima que en el futuro, el modelado por homología será aplicable a la mayoría de las secuencias. Su precisión general cubre un amplio espectro, por lo que los modelos obtenidos pueden tener aplicaciones muy diversas.

1.5.3. Etapas en el modelado por homología

Dada una secuencia problema cuya estructura se quiere predecir, en la práctica todos los protocolos de modelado por homología aplican el esquema de la Figura 1. Si es necesario, los pasos se repiten de manera iterativa hasta que el modelo final es satisfactorio. Aunque ello puede consumir bastante tiempo, puede mejorar la calidad del modelo resultante para casos difíciles (John and Sali, 2003).

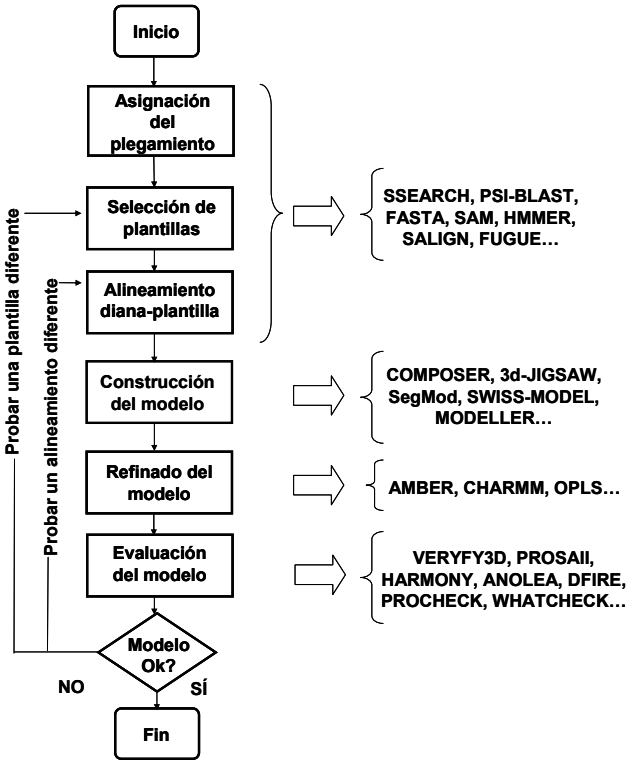


Figura 1. Etapas en el modelado por homología, con ejemplos de programas para llevarlas a cabo.

1.) Asignación del plegamiento y alineamiento diana-plantilla

Aunque éstos son pasos distintos en el proceso de modelado por homología, en la práctica casi todos los métodos de asignación del plegamiento también proporcionan alineamientos secuencia-estructura. La identificación de posibles plantillas se lleva a cabo buscando en las bases de datos de estructuras, proteínas con secuencias homólogas a la de la proteína diana mediante un alineamiento. La similitud detectada usualmente se cuantifica en términos de medidas estadísticas, como el *E-valor*, el *p-valor* o el *z-score*, dependiendo del método usado. A la hora de comparar secuencias, hay que tener en cuenta los siguientes aspectos fundamentales: el tipo de alineamiento, el algoritmo utilizado para encontrar el alineamiento óptimo (el de mejor puntuación), el sistema de puntuación que se usa para valorarlo y el método estadístico para evaluar la significancia de la puntuación dada al alineamiento.

Algoritmos de alineamiento

Dadas dos secuencias, existen numerosos alineamientos posibles entre ellas. Sin embargo, sólo uno representa correctamente su relación de homología. Usualmente no hay forma de saber los eventos ocurridos durante la divergencia de secuencias a lo largo de la evolución, por lo que la única opción es seleccionar como óptimo aquel alineamiento con la puntuación más alta.

Dado que la complejidad del problema de encontrar el alineamiento óptimo es del orden de *NxM* (siendo *N* y *M* las longitudes de las secuencias a alinear), construir todos los posibles alineamientos y puntuarlos es una tarea impracticable para secuencias largas, y es entonces cuando resulta de gran utilidad acudir a la técnica conocida como "*programación dinámica*" en ciencias de la computación, ya que permite simplificar la tarea reduciendo el espacio computacional con un buen compromiso de sensibilidad. La estrategia utilizada es la conocida "*divide y vencerás*". De esta manera, el problema global se subdivide en muchos pequeños problemas que son más fáciles de resolver de forma óptima y una vez resueltos, la solución óptima de cada uno de ellos se combina con la de los demás para dar la solución óptima del problema global. Se trata de un algoritmo recursivo en donde las partes del problema poseen una dependencia secuencial, de manera que para resolver una parte, debe haberse resuelto previamente la inmediatamente anterior.

La idea es subdividir el problema de la comparación de las dos secuencias en subproblemas de comparación de pares individuales de aminoácidos. Las secuencias se alinean en una matriz y se sigue una técnica especial para generar sus elementos, ya que en cada posición del alineamiento se deben considerar no sólo las probabilidades de sustitución de un aminoácido por otro, sino que se ha de tener en cuenta la probabilidad de que haya huecos (inserción o eliminación de residuos). En el caso de dos secuencias, para cada celda de la matriz hay tres posibilidades: hacer un hueco en la secuencia 1, un hueco en la secuencia 2, o una sustitución por el correspondiente aminoácido. En la celda se coloca el mayor valor posible de la puntuación de las tres posibilidades y así se va llenando la matriz. Después se realiza un rastreo inverso de la matriz rellena para obtener el alineamiento óptimo.

Básicamente, existen dos modalidades de alineamiento entre dos secuencias: **alineamiento global** y **alineamiento local**. Se trata de dos variantes de programación dinámica, conocidas respectivamente como algoritmo Needleman-Wunsch (Needleman and Wunsch, 1970) para el alineamiento global y el algoritmo Smith-Waterman (Smith and Waterman, 1981) para el alineamiento local. El primero es apropiado para el alineamiento de dominios independientes. Sin embargo, es muy común que las secuencias muestren sólo regiones locales de similitud (p.ej., múltiples dominios o "*repeats*") que no se podrían detectar en un alineamiento global. Así, el algoritmo de Smith-Waterman se usa en el caso en donde las secuencias a alinear son de muy diferente tamaño y donde resulta ilógico tratar de extender el alineamiento de manera que abarque la longitud de la secuencia más larga. También, es útil para evidenciar zonas acotadas de alta similitud, algo bastante común en las secuencias proteínicas debido a la forma en la que éstas evolucionan (barajado de dominios estructurales). En ese contexto, un alineamiento local se restringe a zonas de alta similitud sin importar las regiones no similares de las secuencias alineadas.

Métodos heurísticos de búsqueda

Los métodos de programación dinámica para alineamientos locales o globales de secuencias, garantizan encontrar la solución óptima y pueden implementarse de manera eficiente. Sin embargo, estos métodos requieren demasiado tiempo como para ser prácticos a la hora de hacer búsquedas contra bases de datos grandes. Para resolver este problema se han desarrollado otras estrategias: son los métodos heurísticos de alineamiento. Éstos no garantizan encontrar la solución óptima, pero en la práctica, raramente pierden una coincidencia particularmente significativa. Generalmente trabajan identificando regiones de interés potencial usando métodos rápidos de registro y expandiendo localmente estas regiones para identificar el mejor alineamiento. Los métodos heurísticos más famosos para alineamiento de secuencias son FASTA (Pearson, 1994) y BLAST (Altschul et al., 1997), con todas sus variantes.

Puntuación y significancia estadística de los alineamientos

Para dar una puntuación a un alineamiento dado, se han desarrollado muchos sistemas diferentes en los que cada posible sustitución de un aminoácido por otro se puntúa de manera independiente, dando lugar a las matrices de sustitución que almacenan esta información (por ejemplo las matrices PAM (Dayhoff et al., 1983), BLOSUM (Henikoff and Henikoff, 1992), etc). Además, también es necesario saber si la puntuación dada es suficientemente alta como para proporcionar una evidencia de homología entre las secuencias alineadas. Para abordar esta cuestión, es útil estimar cómo de alta se espera que sea la puntuación de un alineamiento hecho de manera aleatoria. No hay ninguna teoría matemática para describir la distribución esperada de puntuaciones para alineamientos globales, pero pueden obtenerse fácilmente estimaciones mediante simulaciones comparando la puntuación del alineamiento observado con la de aquellos alineamientos hechos a partir de secuencias aleatorias de la misma longitud y composición que las que se están estudiando (Altschul and Erickson, 1985; Fitch and Smith, 1983).

Sin embargo, sí existe un modelo estadístico propuesto por Karlin y Altschul que proporciona una teoría matemática para describir la distribución esperada de puntuaciones de alineamientos locales aleatorios (Karlin and Altschul, 1990). La forma de la función de densidad de probabilidad obtenida se conoce como la "distribución de valores extremos" (*extreme value distribution*). Relacionando una puntuación de alineamiento observada *S*, con la distribución esperada y el número de secuencias contenidas en la base de datos, es posible calcular analíticamente la significancia estadística en la forma de un *E-valor*. La interpretación simple de un *E-valor* es el número de alineamientos con puntuaciones de al menos *S*, que se podrían esperar únicamente por azar. La significancia de un alineamiento también depende del tamaño del espacio de búsqueda que se ha usado, ya que en bases de datos más grandes se dan más alineamientos al azar.

Métodos de alineamiento

Las relaciones secuencia-estructura se pueden clasificar de manera general en tres categorías diferentes: (i) relaciones fácilmente detectables, caracterizadas por una identidad en secuencia > 30 %; (ii) la "zona en penumbra" (*twilight zone*) (Rost, 1999), que corresponde a relaciones estadísticamente significativas (es decir, *E-valores* altos), pero caracterizadas por estar entre el 10-30% de identidad en secuencia; y (iii) la "zona de medianoche" (*midnight zone*) (Rost, 1999), que corresponde a similitudes en secuencia no significativas, es decir, no detectables con los métodos actuales en ausencia de estructura. Dependiendo de la zona del espectro de identidad en secuencia en que se encuentren la diana y la plantilla, se pueden utilizar diferentes métodos para construir el alineamiento entre ellas:

Métodos secuencia-secuencia: Para secuencias proteínicas estrechamente relacionadas con identidades mayores del 30-40%, los alineamientos producidos por todos los métodos son casi siempre correctos en las regiones conservadas. La manera más rápida para buscar posibles plantillas en este régimen de identidad es usar métodos de alineamiento de pares de secuencias como SSEARCH (Pearson, 1994), aunque también pueden emplearse BLAST (Altschul et al., 1997) y FASTA (Pearson, 1994). Sin embargo, Brenner et al. mostraron que estos métodos detectan solamente ~18% de los pares homólogos por debajo del 40% de identidad, mientras que son capaces de identificar más del 90% de las relaciones cuando la identidad en secuencia está entre el 30-40% (Brenner et al., 1998). Otra prueba, basada en 200 alineamientos estructurales de referencia con porcentajes de identidad en secuencia entre el 0-40% indicó que BLAST alinea correctamente sólo el 26% de las posiciones (Sauder et al., 2000). Por ello, siempre que sea posible, es preferible utilizar alineamientos secuencia-perfil o perfil-perfil.

Métodos secuencia-perfil: A medida que las relaciones en secuencia se desplazan hacia la zona de penumbra, resulta más difícil obtener una buena sensitividad en la búsqueda y en la calidad del alineamiento (Rost, 1999; Saqi et al., 1998). La introducción de los métodos de perfiles por Gribskov et al (Gribskov et al., 1987), supuso una mejora significativa. El **perfil** de una secuencia se deriva de alineamientos múltiples de secuencia y especifica tipos de residuo para cada posición del alineamiento. La información en un alineamiento múltiple de secuencias a menudo se codifica en forma de matriz de puntuación específica de posición (*position specific scoring matrix*, PSSM) (Altschul et al., 1997; Henikoff and Henikoff, 1996; Henikoff and Henikoff, 1994), o como un modelo oculto de Markov (*hidden markov model*, HMM) (Eddy, 1998; Krogh et al., 1994). Para poder identificar posibles plantillas para el modelado por homología, el perfil de la secuencia diana se usa para buscar en las bases de datos de secuencias de plantillas. Los métodos perfil-secuencia tienen más sensitividad para detectar estructuras relacionadas en la zona de penumbra que los métodos basados en pares de secuencias; detectan aproximadamente el doble de homólogos que éstos por debajo del 40% de identidad en secuencia (Lindahl and Elofsson, 2000). Además, los alineamientos perfil-secuencia resultantes

alinean correctamente aproximadamente el 43-48% de los residuos en el rango de 0-40% de identidad en secuencia (Marti-Renom et al., 2004), lo que es casi el doble que los métodos de pares de secuencias. Los programas más usados para alineamientos perfil-secuencia son PSI-BLAST (Altschul et al., 1997), SAM (Karplus et al., 1998), HMMER (Eddy, 1998) y BUILD_PROFILE (Eswar, 2005).

Métodos perfil-perfil: Estos métodos aparecen como una extensión natural de los anteriores. Utilizan el perfil de la secuencia diana para buscar posibles estructuras plantilla en una base de datos de perfiles de plantillas. Se ha demostrado que estos métodos producen los mejores resultados en esta etapa del modelado (Marti-Renom et al., 2004). Los métodos perfil-perfil detectan ~28% más de relaciones que los anteriores a nivel de superfamilia y mejoran la calidad del alineamiento en un 15-20% comparado con éstos(Marti-Renom et al., 2004). Hay un gran número de variantes en los métodos perfil-perfil dependiendo de las funciones de puntuación que usen (Marti-Renom et al., 2004). Sin embargo, se ha demostrado que el comportamiento general de todas ellas es comparable (Marti-Renom et al., 2004). Algunos de los programas que implementan este tipo de métodos son FFAS (Jaroszewski et al., 2005), SP3 (Zhou and Zhou, 2005), SALIGN (Marti-Renom et al., 2004) y PPSCAN (Eswar, 2005).

Métodos de hilado secuencia-estructura: Los métodos de hilado secuencia-estructura suponen una alternativa a los alineamientos perfil-perfil, aunque se ha demostrado que en general, estos últimos son superiores hoy en día. Estos métodos consiguen mayor sensitividad usando información estructural derivada de las plantillas. La calidad del ajuste secuencia-estructura se mide en función de potenciales de compatibilidad secuencia-estructura (Godzik, 2003). Los esquemas de puntuación usados se basan tanto en tablas de sustitución de residuos dependientes de características estructurales (exposición al solvente, tipo de estructura secundaria, y propiedades de enlace de hidrógeno (Shi et al., 2001; Zhou and Zhou, 2005)), como en potenciales estadísticos para las interacciones de residuos derivados del alineamiento (Bowie et al., 1991; Sippl, 1995; Skolnick and Kihara, 2001). Programas usados comúnmente son GenTHREADER (McGuffin and Jones, 2003), 3D-PSSM (Kelley et al., 2000), FUGUE (Shi et al., 2001), SP3 (Zhou and Zhou, 2005) y SAM-T02 multi-track HMM (Karplus et al., 2003).

Una vez obtenida una lista de estructuras de proteínas y sus alineamientos con la secuencia diana, las estructuras plantilla se ordenan dependiendo del propósito del modelado. Éstas, se pueden escoger basándose puramente en la identidad en secuencia entre la diana y la plantilla, o en una combinación de otros criterios, como calidad experimental de las estructuras, conservación de residuos del sitio activo, holo-estructuras con ligandos de interés unidos, e información biológica previa sobre el solvente, pH y contactos cuaternarios. No es necesario seleccionar una única plantilla. De hecho, el uso de varias de ellas aproximadamente

equidistantes de la secuencia diana, generalmente aumenta la calidad del modelo (Srinivasan and Blundell, 1993).

2). Construcción del modelo

A partir del alineamiento diana-plantilla, se pueden usar diferentes métodos para construir un modelo 3D de la proteína diana. Para ello, básicamente existen tres técnicas diferentes: ensamblado de cuerpos rígidos (*rigid-body assembly*) (Blundell et al., 1987), utilización de segmentos coincidentes (*segment matching*) (Bystroff and Baker, 1998), o satisfacción de restricciones espaciales (*spatial restraints*) (Sali and Blundell, 1993). El primer método construye el modelo a partir de unas pocas regiones centrales, y a partir de lazos y cadenas laterales obtenidos de estructuras relacionadas. Ejemplos de programas que implementan este método son COMPOSER (Nagarajaram et al., 1999), 3D-JIGSAW (Bates et al., 2001), y SWISS-MODEL (Schwede et al., 2003). El segundo, se basa en las posiciones aproximadas de átomos conservados de las plantillas para calcular las coordenadas de otros átomos. Un ejemplo de programa que utiliza este método es SegMod (Levitt, 1992). El tercer grupo de métodos, usa tanto geometría de distancias como técnicas de optimización (fundamentalmente dinámica molecular restringida), para satisfacer restricciones espaciales obtenidas a partir del alineamiento de la secuencia diana con las estructuras plantilla. Un ejemplo de programa que implementa este método es MODELLER (Fiser et al., 2002).

En cuanto al **modelado de las regiones divergentes** o lazos, básicamente hay dos clases de métodos: aproximaciones de búsqueda en bases de datos de estructuras para encontrar segmentos que se ajusten a las regiones de anclaje (Chothia and Lesk, 1987; Jones and Thirup, 1986) y aproximaciones de búsqueda conformacional que se apoyan en la optimización de una función de puntuación (Bruccoleri and Karplus, 1987; Moult and James, 1986). También existen métodos mixtos que combinan ambas (Deane and Blundell, 2001; van Vlijmen and Karplus, 1997). Ejemplos de bases de datos de lazos y de programas para modelarlos son Sloops (Burke et al., 2000), ArchDB (Espadaler et al., 2004) y ModLoop (Fiser and Sali, 2003).

Las **cadenas laterales** a menudo se modelan utilizando librerías de rotámeros, construidas a partir del hecho de que en las estructuras cristalográficas de alta resolución, la mayoría de ellas se pueden representar mediante un número limitado de confórmeros que cumplen una serie de restricciones estereoquímicas y energéticas. Un ejemplo de programa para modelar cadenas laterales es SCWRL (Canutescu et al., 2003).

3). Refinado del modelo

Una vez construido el modelo inicial, el siguiente paso es preguntarse si su estructura se puede mejorar. El refinado es una tarea difícil que requiere una estrategia de muestreo efectiva así como una función de energía de calidad para guiar la búsqueda a través del espacio

conformacional. Algunos intentos de refinado consisten en minimizar su energía o en aplicar dinámica molecular. Normalmente, este paso consiste en aplicar minimizaciones cortas (de unos 100-1000 pasos) de descenso más pronunciado (*steepest-descent*) o gradiente conjugado, hasta convergencia, con el objetivo de eliminar choques entre átomos. Lee et al. usaron simulaciones de dinámica molecular con Poisson-Boltzman (MM-PBSA) en dos proteínas pequeñas, HP-36 y S15, y mostraron que las estructuras nativas se podían distinguir de los modelos de baja resolución y que el estado nativo era estable (Lee et al., 2001a), pero para proteínas grandes no lograron obtener mejores estructuras (Lee et al., 2001b). Sin embargo, la combinación de restricciones locales con potenciales basados en conocimiento y aproximaciones de dinámica molecular conseguía algunas mejoras sobre los estudios que sólo usaban métodos de dinámica molecular (Lu and Skolnick, 2003). Otra aproximación consistió en el uso de múltiples plantillas (Contreras-Moreira et al., 2003) y la aplicación de un conjunto de movimientos en el espacio conformacional (Offman et al., 2006) para tratar de mejorar la calidad final. Sin embargo, los resultados obtenidos con todos estos métodos todavía no son lo suficientemente satisfactorios y la mayoría de las veces, su aplicación conduce a la degradación del modelo en vez de a la mejora de su calidad. No obstante, recientemente se ha demostrado que el uso de protocolos que combinan el refinado basado en energía con el muestreo a lo largo de direcciones favorecidas evolutivamente (Qian et al., 2004) y con el uso de restricciones derivadas de estructuras homólogas (Misura et al., 2006) consiguen obtener de manera consistente modelos de mayor calidad que las plantillas de partida y presentan un camino muy prometedor para tratar de mejorar esta etapa del modelado por homología.

4). Detección de errores en los modelos

El primer paso es determinar si el modelo obtenido tiene el plegamiento correcto. Hay muchos métodos que usan perfiles 3D y potenciales estadísticos (Luthy et al., 1992; Melo et al., 2002; Sippl, 1990) que evalúan la compatibilidad entre la secuencia y la estructura modelada, comparando el entorno de cada residuo en el modelo, con respecto al entorno esperado encontrado en las estructuras de rayos-X de alta resolución. Ejemplos de programas que incorporan estos métodos son VERIFY3D (Luthy et al., 1992; Melo et al., 2002; Sippl, 1990), PROSAII (Sippl, 1993), HARMONY (Topham et al., 1994), ANOLEA (Melo and Feytmans, 1998) y DFIRE (Zhou and Zhou, 2002). Una buena estrategia para caracterizar regiones erróneas en los modelos consiste en usar diferentes métodos y tratar de identificar un consenso entre todos ellos.

El modelo debe someterse también a evaluaciones de autoconsistencia para asegurar la satisfacción de las restricciones usadas para obtenerlo. Un requerimiento básico para un buen modelo es que tenga una buena estereoquímica (longitudes y ángulos de enlace, ángulos torsionales de la cadena principal, mapa de Ramachandran y contactos no enlazantes

principalmente). Los programas más útiles para evaluar la estereoquímica son PROCHECK (Laskowski et al., 1993) y WHATCHECK (Hooft et al., 1996). Además de una buena estereoquímica, un buen modelo también ha de tener una baja energía, de acuerdo a campos de fuerza de mecánica molecular, como CHARMM (Brooks et al., 1983), por ejemplo. Sin embargo, una baja energía, no asegura un modelo correcto.

1.5.4. Fuentes de error en el modelado por homología

Las fuentes de error más importantes en los modelos obtenidos mediante esta técnica son las siguientes (Marti-Renom et al., 2000):

- Selección de plantillas incorrectas. Éste es un problema potencial cuando se usan como modelos proteínas estructuralmente muy distantes de la diana. Distinguir entre un modelo basado en una plantilla incorrecta y uno basado en un alineamiento incorrecto con una plantilla correcta es muy difícil. En ambos casos, los métodos de evaluación predirán un modelo erróneo. La conservación de los residuos clave funcionales o estructurales en la secuencia problema aumenta la confianza en la asignación de un plegamiento dado.
- Alineamientos incorrectos. Es la fuente de error más frecuente, especialmente cuando la identidad en secuencia entre la proteína diana y la plantilla decae por debajo del 30 %. Es muy importante obtener el alineamiento más preciso posible, porque ningún método actual permite generar un modelo correcto a partir de un alineamiento incorrecto. Sin embargo, este tipo de error se puede minimizar de dos maneras. Primero, utilizando las técnicas secuencia-perfil y perfil-perfil descritas anteriormente. La segunda manera de mejorar el alineamiento es modificando iterativamente aquellas regiones del mismo que corresponden a posibles errores previamente detectados en el modelo (Sanchez and Sali, 1997).

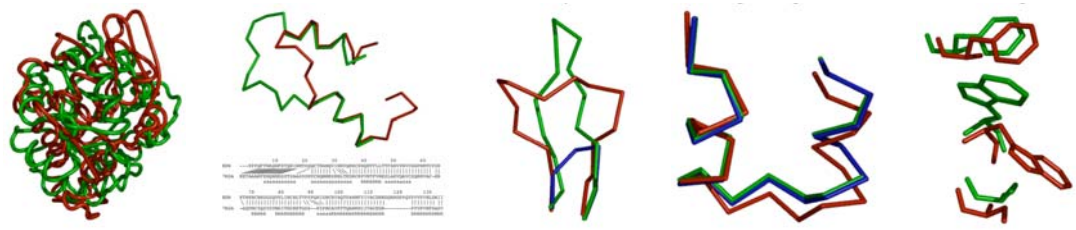


Figura 2. Errores típicos en el modelado por homología. Se muestran las fuentes típicas de error encontradas en los modelos generados con esta metodología. La diana a modelar se representa en color rojo; la plantilla utilizada en azul y el modelo resultante en verde. (Adaptado de (Marti-Renom et al., 2000)).

- Regiones estructurales que no existen en la plantilla. Los segmentos de la secuencia diana que no tienen región equivalente en la estructura de la plantilla (por ejemplo, inserciones o lazos), son las regiones más difíciles de modelar. De nuevo, cuando la relación entre la diana y la plantilla es muy distante, los errores en el alineamiento pueden conducir a posiciones incorrectas de las inserciones. Usar métodos de alineamiento que incorporan información estructural, a menudo puede corregir estos errores. Una vez obtenido un alineamiento fiable, las conformaciones de los lazos, para inserciones de menos de 8-10 residuos se pueden predecir correctamente con varios protocolos de modelado (Fiser and Sali, 2003; Jacobson et al., 2004; van Vlijmen and Karplus, 1997).
- Distorsiones y desplazamientos en regiones correctamente alineadas. Como consecuencia de la divergencia en secuencia, la conformación de la cadena principal cambia incluso si el empaquetamiento global se mantiene. Así, es posible que en algunos segmentos correctamente alineados de un modelo, la plantilla sea localmente diferente (<3 Å) de la diana, dando errores en esa región. Las diferencias estructurales muchas veces no se deben a diferencias en secuencia, sino que son consecuencia de artefactos en la determinación estructural en entornos diferentes. El uso simultáneo de muchas plantillas puede minimizar este tipo de error (Sanchez and Sali, 1997; Srinivasan and Blundell, 1993).
- Empaquetamiento de las cadenas laterales incorrecto. A medida que las secuencias divergen, el empaquetamiento de las cadenas laterales cambia. Algunas veces, incluso siendo idénticas, puede que la conformación de las cadenas laterales no se conserve, lo que supone un escollo más en estos métodos. Los errores en las cadenas laterales son críticos si ocurren en regiones involucradas en la función de la proteína, como los sitios activos y de unión a ligando.

1.5.5. Aplicaciones de los modelos obtenidos por homología

Dependiendo de la similitud en secuencia entre la diana y la plantilla, la calidad de los modelos construidos con esta metodología, y, por tanto, sus posibles aplicaciones, son muy variadas (Blundell and Johnson, 1993; Gao et al., 2003; Thiel, 2004) (Figura 3): diseñar mutantes específicos para probar hipótesis acerca de la función de una proteína; identificar centros activos y sitios de unión a ligando; buscar, diseñar y mejorar ligandos para un sitio de unión dado; modelar la especificidad al sustrato; diseñar fármacos; predecir interacciones proteína-proteína; inferir función; facilitar cálculos de reemplazamiento molecular en determinación estructural por R-X; refinar modelos basados en restricciones de RMN; confirmar relaciones estructurales remotas entre proteínas; evaluar y mejorar alineamientos secuencia-estructura; facilitar el análisis estructural y la reconstrucción tridimensional de complejos y orgánulos celulares mediante combinación de los modelos obtenidos con técnicas

de crio-microscopía electrónica; racionalizar observaciones experimentales conocidas y diseñar nuevos experimentos.

Afortunadamente, no es necesario que un modelo tridimensional sea absolutamente perfecto para ser útil en biología. Sin embargo, el tipo de problema que se puede pretender resolver con un modelo particular depende de su calidad. En el extremo inferior del espectro de calidad, están los modelos basados en menos del 25% de identidad en secuencia entre la diana y la plantilla, que suelen tener menos del 50% de sus Cα a una distancia de menos de 3.5 Å respecto a sus posiciones correctas. Estos modelos mantienen un plegamiento correcto y siguen siendo de utilidad, puesto que con frecuencia es suficiente conocer únicamente el plegamiento de una proteína para acotar su posible función bioquímica a un pequeño número de posibilidades. En la mitad del espectro de calidad están los modelos basados en aproximadamente el 35 % de identidad en secuencia, lo que a menudo corresponde a tener el 85% de los Cα modelados a una distancia menor de 3.5 Å respecto a sus posiciones correctas. Afortunadamente, los centros activos y de unión a ligando a menudo están más conservados que el resto de la proteína y por tanto, pueden ser modelados con más exactitud. En general, los modelos de resolución media resultan útiles para refinar las predicciones funcionales basadas en secuencia únicamente, ya que la unión a ligando está más determinada directamente por la estructura del sitio de unión que por su secuencia. A menudo con estos modelos es posible predecir características de la proteína diana que no están presentes en la estructura plantilla. Los modelos de resolución media también se pueden usar para construir mutantes específicos con capacidad de unión alterada o destruida, que se pueden utilizar para probar hipótesis sobre relaciones secuencia-estructura-función. Así mismo, se puede abordar también el diseño de construcciones para mejorar la cristalización.

El extremo superior del espectro de calidad corresponde a modelos basados en el 50% de identidad en secuencia o más. La calidad media de estos modelos es comparable a la de las estructuras cristalográficas de baja resolución (unos 3 Å), o estructuras de RMN de resolución media (10 restricciones de distancia por residuo). Los alineamientos en los que se basan estos modelos generalmente casi no contienen errores. Además de las aplicaciones ya mencionadas, los modelos de alta calidad se pueden usar para anclaje de pequeños ligandos o de proteínas enteras contra otras proteínas. Mención especial merece también la aplicación de estos modelos al diseño de fármacos, puesto que son muy útiles en la detección de posibles ligandos mediante el uso de técnicas de cribado virtual, así como en la búsqueda de los sitios que confieran selectividad dentro de las familias de proteínas.

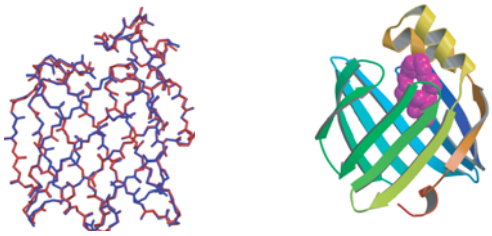
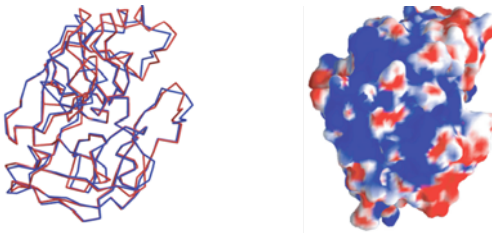
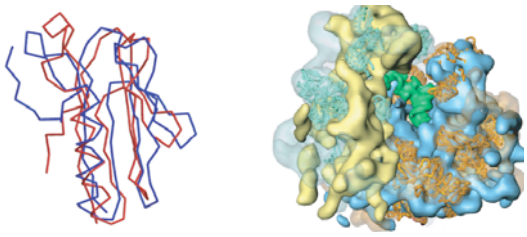
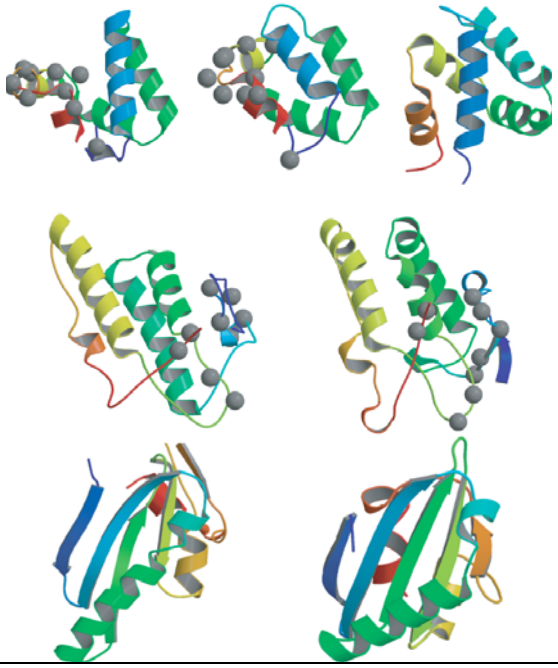
RMN, R-X	**Porcentaje de identidad en secuencia**	**100**	**Precisión del modelo**	1.0 ≈ 100%		**APLICACIONES** - Estudiar mecanismos catalíticos - Diseñar y mejorar ligandos - Docking de macromoléculas
MODELADO POR HOMOLOGÍA		**50**		1.5 ≈ 95%		- Cribado virtual y docking de pequeños ligandos - Reemplazamiento molecular en R-X
RECONOCIMIENTO DE PLEGAMIENTO		**30**		3.5 ≈ 80%		- Diseño de quimeras estables y cristalizables - Proponer mutagénesis dirigidas - Refinar estructuras de RMN
PREDICCIÓN DE NOVO		**Similitud insignificante**		4.0 ≈ 80aa		- Ajuste en los mapas de densidad electrónica de baja resolución - Encontrar sitios funcionales mediante búsqueda de motivos 3D - Estructura a partir de unas pocas restricciones experimentales - Anotación de función mediante asignación de plegamiento - Establecer relaciones evolutivas

Figura 3. Calidad y aplicaciones de modelos estructurales de proteínas. Se muestran los diferentes rangos de aplicabilidad del modelado por homología, el reconocimiento de plegamiento y de la predicción estructural *ab initio*; su calidad correspondiente y ejemplos de aplicaciones. (Tomado de (Baker and Sali, 2001)).

1.5.6. Evaluación de los protocolos de modelado por homología

Dada la abundante oferta de métodos de predicción disponibles, es indispensable, tanto para los desarrolladores de los mismos como para sus usuarios, conocer el grado de precisión y calidad que tienen. Para evaluarlo, existe una serie de iniciativas que permiten medir el progreso experimentado por las diferentes técnicas, establecer el estado actual de las metodologías y señalar las áreas en las que se deben centrar todos los esfuerzos para hacer avanzar el campo. Así, se han desarrollado experimentos como CASP (Venclovas et al., 2001), CAFASP (Fischer et al., 2001), LiveBench (Bujnicki et al., 2001) y EVA (Koh et al., 2003).

1.5.7. Limitaciones y retos actuales. Lecciones aprendidas en CASP

CASP (*Critical Assessment of Techniques for Protein Structure Prediction*), es un experimento (aunque normalmente se considera como una competición), abierto a toda la comunidad científica, cuya primera edición tuvo lugar en 1994-95 y que se organiza cada dos años(CASP, 1994). En él, se invita a los participantes a predecir las estructuras tridimensionales de una serie de proteínas a partir de sus secuencias de aminoácidos. Las estructuras de estas proteínas ya han sido resueltas experimentalmente a su vez mediante cristalografía de rayos X o RMN, pero no están disponibles en el momento de la competición, y por tanto, son completamente desconocidas para los grupos participantes.

Los resultados de los experimentos de CASP se publican al final de cada competición. En el apartado de modelado por homología, en todas las ediciones celebradas, incluyendo los últimos CASP5 y CASP6 (actualmente está en fase de desarrollo el CASP7), se ha puesto de manifiesto que la calidad del modelo depende crucialmente de la calidad del alineamiento diana-plantilla a partir del cual se construye (Tramontano and Morea, 2003). La técnica se ha beneficiado del continuo incremento de secuencias y estructuras resueltas experimentalmente (cuantas más estructuras se conocen, más posibilidades hay de encontrar un homólogo estructural cercano a la diana), y del desarrollo de nuevos algoritmos capaces de detectar relaciones evolutivas más distantes en secuencia. Sin embargo, la mejora en la calidad de los alineamientos secuencia-estructura constatado en las sucesivas ediciones de CASP no parece traducirse en una mejora obvia en la calidad de los modelos tridimensionales finales. Es más, incluso en ausencia de errores en los alineamientos y usando las plantillas óptimas, se ha demostrado que los métodos de modelado basados en patrones tienen limitaciones intrínsecas (Contreras-Moreira et al., 2005). La característica que resume estos problemas es que la mayoría de los modelos obtenidos carecen de una estructura significativamente más cercana a la de la diana que la del mejor patrón utilizado, y que en los pocos casos donde una pequeña mejora se obtuvo, ésta nunca suele ser mayor de 0.4 Å (Tramontano and Morea, 2003). Se piensa que las causas de error en el refinado son principalmente las inexactitudes en los campos de fuerza

actuales (que no pueden simular con la suficiente precisión el delicado balance de fuerzas en el estado nativo de la proteína, con muchas interacciones), y las dificultades que supone el muestreo de un gran número de conformaciones alternativas en la búsqueda del mínimo global (Qian et al., 2004).

El refinado es particularmente relevante cuando la identidad en secuencia entre la diana y la plantilla está por debajo del 30%. Los modelos en este rango construidos usando los métodos actuales generalmente tienen un RMSD > 1.5-2.0 Å en la región estructuralmente conservada, lo que corresponde también al caso más frecuente en un proyecto de modelado. Se ha estimado que la probabilidad de que la secuencia problema comparta menos del 30% de identidad con alguna proteína de estructura conocida con el mismo plegamiento es de al menos el 50%(Marti-Renom et al., 2000), por lo que el refinado de los modelos es un problema importante, tanto, que se plantea como un nuevo apartado independiente a evaluar en la actual edición de CASP7 (Tress et al., 2005).

1.6. Hipótesis de trabajo

Se prevé que el refinado estructural de los modelos de proteínas a partir de la identificación de un alineamiento secuencia-estructura puede llegar a ser pronto un cuello de botella importante a la hora de hacer estos métodos de predicción útiles para la biología.

Los problemas existentes a la hora de llevarlo a cabo podrían provenir en parte, del desconocimiento que se tiene acerca de cómo cambian las estructuras a lo largo de la evolución.

La comparación de estructuras conocidas pertenecientes a una familia de proteínas homólogas, y la caracterización desde un punto de vista físico de las deformaciones estructurales que han tenido lugar dentro de esa familia, podrían ayudar a adquirir este conocimiento.

Estudiar el proceso de deformación estructural en detalle y poner las herramientas y protocolos a punto para su aplicación al modelado estructural, podría conducir a la obtención de modelos de proteínas de mayor calidad.

OBJETIVOS

2. Objetivos

En esta tesis se estudia el proceso de evolución estructural en proteínas homólogas mediante el análisis de sus patrones de deformación. Se explora la utilización de estos patrones para ayudar a resolver deficiencias de los métodos actuales de predicción estructural relacionadas con la etapa de refinado, con el propósito final de lograr la obtención de modelos de mayor calidad. Para cumplir este objetivo global se abordan los siguientes objetivos parciales:

- **Desarrollar un nuevo algoritmo de alineamiento estructural múltiple para la caracterización de las regiones estructuralmente conservadas de las familias de proteínas homólogas y la determinación de las distorsiones estructurales evolutivas.** Se pretende desarrollar un algoritmo determinista, pero rápido, adaptado a estudios a gran escala, capaz de dar alineamientos de gran calidad y extensos centros (*cores*) que permitan caracterizar las deformaciones estructurales en familias de proteínas.

- **Estudiar la relación entre las deformaciones estructurales observadas en la evolución y las propiedades dinámicas de la topología de la proteína.** Entender y caracterizar desde un punto de vista físico estos cambios estructurales puede ayudar a definir un espacio de baja dimensionalidad en el que se representen satisfactoriamente las estructuras nativas.

- **Utilizar esta información para mejorar el muestreo conformacional en las técnicas de modelado por homología.** El uso de un espacio de baja dimensionalidad puede facilitar el muestreo en la búsqueda de posibles modelos para una determinada proteína diana.

MATERIALES Y MÉTODOS

3. Materiales y Métodos

3.1. Conjuntos de datos

Cada estudio realizado en esta tesis, necesitó la preparación de un conjunto de datos adaptado y diferente. A continuación, se detalla cada uno por separado:

3.1.1. Conjuntos de datos de entrenamiento y evaluación para el desarrollo del algoritmo de alineamiento estructural múltiple

Se utilizaron cinco conjuntos de datos de:

(1). HOMSTRAD. Se utilizó un conjunto de 105 alineamientos estructurales de la base de datos de HOMSTRAD (una base de datos de alineamientos estructurales múltiples para familias homólogas) (Mizuguchi et al., 1998b). Estos alineamientos se utilizan como referencia (*gold standard*) para el nivel de familia, ya que están corregidos a mano.

(2). CAMPASS. Este conjunto consta de 551 alineamientos manuales al nivel de superfamilia (Sowdhamini et al., 1998), derivados de manera similar a los de HOMSTRAD. Se adoptan como referencia para el nivel de superfamilia.

(3). Multiprot. Conjunto de estructuras correspondientes al conjunto CAMPASS descrito arriba, pero con los alineamientos generados usando el programa MultiProt (Shatsky et al., 2004). Los alineamientos de secuencia correspondientes a los de estructura generados, se obtienen aplicando el programa Stacatto (M. Shatsky, comunicación personal) a la salida de Multiprot.

(4). FSSP. FSSP es una colección de alineamientos estructurales generados con el programa Dali. Aunque estrictamente, no son alineamientos estructurales múltiples, se han incluido porque la base de datos FSSP es una fuente muy popular de comparaciones múltiples de estructura. Para construir este conjunto, se llevó a cabo una comparación de todos-contra-todos con MAMMOTH de pares sobre un conjunto de alineamientos FSSP (Holm and Sander, 1998). Se usó un algoritmo de detección de cliqué para seleccionar conjuntos de estructuras que pudieran ser alineadas tanto con MAMMOTH-mult como con FSSP. Se obtuvieron 1385 cliqués, donde todos los miembros del cliqué estaban por encima de un cierto valor de corte para la similitud estructural, de acuerdo con MAMMOTH y FSSP. Los resultados de aplicar MAMMOTH-mult a este conjunto de datos se compararon con los alineamientos procesados por DaliLite (Holm and Park, 2000), usando unos parámetros de calidad descritos en la parte de métodos.

(5). Superplegamientos. Éste es un conjunto formado por los alineamientos de proteínas pertenecientes a dos tipos diferentes de superplegamientos (inmunoglobulinas y globinas), cuyos alineamientos estructurales y clasificación han sido estudiados en detalle por diferentes autores. El conjunto de inmunoglobulinas está formado por 26 dominios analizados por Bork et al. (Bork et al., 1994), que clasificó manualmente estas estructuras en cuatro grupos distintos.

También se estudió el plegamiento de las globinas porque es un plegamiento clásico analizado en muchas investigaciones sobre alineamientos estructurales múltiples. Como referencia para este conjunto de datos se usó la clasificación de SCOP.

3.1.2. Conjunto de datos para el estudio de la relación entre los movimientos estructurales evolutivos y topológicos de las proteínas

El conjunto de datos usado en este apartado consiste en un conjunto de 35 superfamilias tomadas de ASTRAL40, bien pobladas y estudiadas, representativas de las 5 clases de proteínas más importantes de SCOP (todo-α, todo-β, α/β, α+β y proteínas pequeñas). El porcentaje de identidad máximo entre dos proteínas dentro de cada superfamilia es 40% (ver Tabla-Mat.Sup. 1 para más detalles).

3.1.3. Conjuntos de datos para la mejora del muestreo conformacional en las técnicas de modelado por homología

En este apartado se usaron tres conjuntos de datos. El primero, es básicamente el mismo usado en el apartado anterior, del cual se han excluido las 5 superfamilias pertenecientes a la clase de proteínas pequeñas. Por tanto, el primer conjunto de datos en este apartado está constituido por 30 superfamilias (conjunto de datos 3.1.3.1).

Para comparar nuestros resultados con los obtenidos por los protocolos de modelado por homología normalmente usados por la comunidad científica, se utilizaron dos conjuntos de datos procedentes de CASP5 (Mayo, 2002): por un lado, una colección de 67 dianas utilizadas en dicha competición (conjunto de datos 3.1.3.2), y por otro, una colección de los mejores modelos generados para esas dianas por los grupos participantes (proporcionado por Bruno Contreras-Moreira) (conjunto de datos 3.1.3.3).

Para modelar las 67 dianas utilizadas en CASP5 se construye una librería de estructuras plantilla a partir de la base de datos de PDB-SELECT (Hobohm et al., 1992) al 90% de redundancia, correspondiente a Abril de 2002. Esta librería está compuesta de 6182 cadenas de pdb. Para cada proteína problema se aplican los siguientes pasos:

1). Se obtiene una lista de posibles homólogos (plantillas), definidos como aquellas estructuras cuya comparación con MAMMOTH de pares (Ortiz et al., 2002) con respecto a la proteína problema (diana), tiene una puntuación por encima de 4.

2). Se descartan aquellas plantillas demasiado largas o demasiado cortas (longitudes 2.5 ó 0.4 veces la longitud de la proteína diana).

3). Se selecciona el conjunto óptimo de plantillas de entre todas las obtenidas en el paso anterior. Para ello, se aplica una optimización de Metropolis Monte Carlo (ver Algoritmo 1): (Nota: En el apartado 3.8.1 se detallan los fundamentos teóricos de esta metodología. No obstante, se adelanta aquí una aplicación particular al caso concreto que nos ocupa):

```
pick_structures (Σ, S_old )                    /*Random selection of structures*/
A_old = mammoth_mult (S_old)                   /*A is defined as A = - % core*/
A_best = A_old; S_best = S_old
for T = T_high to T_low
      for n = 1 to nstep
            modify_selection (S_old, S_new)    /*Implements the "move-set"*/
            A_new = mammoth_mult (S_new)
            ΔA = A_new - A_old
            p = random (seed)
            if (p < exp (-ΔA/T)) then          /*Metropolis criterion*/
                  S_old = S_new;
                  A_old = A_new
            endif
            if (A_new < A_best)  A_best = A_new; S_best = S_new /*Keeps best*/
      enddo
enddo
```

Algoritmo 1. Monte Carlo para la selección de plantillas. Ver el texto principal para más detalles. La subrutina *pick_structures* no se describe porque conocer su funcionamiento resulta fácil a partir de los comentarios de la descripción del Algoritmo 1 y su implementación es sencilla.

La constante T desempeña aquí el mismo papel que la temperatura cuando se simulan sistemas físicos. Y los valores de los parámetros del algoritmo se establecieron de manera que se optimizara la relación inicial de aceptación/rechazo y el número de intentos. De esta manera, $n_{step} = 50$, $T_{high} = 1.5$, $T_{low}=0.1$.

Los subconjuntos de estructuras finalmente seleccionados se pueden considerar como aquellos que proporcionan el mayor espacio estructural de muestreo disponible para representar cada diana, en el momento en que tuvo lugar CASP5 (Mayo de 2002). Para todas las dianas se encontraron al menos 3 plantillas y se modeló en promedio más del 60% de la estructura total de las mismas (ver apartado de Resultados, pág. 89).

3.2. Alineamiento estructural múltiple. Algoritmo de MAMMOTH-mult

En el trabajo realizado en esta tesis es crucial disponer de una buena caracterización de los *centros estructurales* de las familias de proteínas homólogas. Los alineamientos múltiples de estructuras permiten dicha caracterización, diferenciando entre regiones conservadas y divergentes. Sin embargo, a pesar de toda la investigación previa en el campo, todavía se requieren algoritmos que tengan un balance adecuado entre calidad y velocidad para aplicaciones a gran escala. Se presenta aquí un nuevo algoritmo, MAMMOTH-mult, para obtener alineamientos múltiples de estructura, que afronta este problema (ver Algoritmo 2). Se trata de un algoritmo determinista, pero rápido, adaptado a estudios a gran escala, capaz de dar alineamientos de gran calidad y extensos centros estructurales, y proporciona el punto de partida

para el análisis de la plasticidad de las familias de proteínas homólogas. El algoritmo se discute a continuación.

```
mammoth-mult
for i = 1 to n-1
      for j = i+1 to n
            S(i,j) = mammoth(x_i,x_j)   /*Mammoth pairwise comparison*/
            S(j,i) = S(i,j)
      enddo
enddo
average_linkage (S, n_c, C)
for n=1 to n_c
      for i = 1 to l_A(n)               /*Align. length of subcluster A in cluster n*/
            for j = 1 to l_B(n)         /*Align. length of subcluster B in cluster n*/
                  S(i,j) = 0
                  for k = 1 to n_A(n)
                        for l = 1 to n_B(n)
                              S(i,j) + URMS(i,j,k,l)
                        enddo
                  enddo
            enddo
      enddo
      needleman-wunsch(S) /*Alignment corresp l_A(n) and l_B(n) are overwritten*/
      maxsub              /*Core is defined here*/
      for i = 1 to l_A (n)
            for j=1 to l_B(n)
                  S(i,j) = 0
                  for k = 1 to n_A(n)
                        for l = 1 to n_B(n)
                          st=w_S*URMS(i,j,k,l)+w_d*exp(-a*d(i,j,k,l)**2
                          S(i,j) = S(i,j) + st
                        enddo
                  enddo
            enddo
      enddo
      needleman-wunsch(S)       /*Reassigment of corresp. obtained above*/
      dowhile (errdif > cutoff)
            errold=err; err = 0
            for i = 1 to (n_A(n) + n_B(n))
                  err = err + simplex(x_i)
            enddo
            errdif = errold - err
      enddo
enddo
```

Algoritmo 2. Pseudocódigo de la sección principal de MAMMOTH-mult. Ver texto principal para más detalles. Las subrutinas average_linkage, needleman_wunsch y simplex no se describen ya que corresponden a algoritmos estándar descritos en (Johnson and Wichern, 1998; Needleman and Wunsch, 1970; Numerical-Recipies-Software, 1992) respectivamente.

El algoritmo (ver también Algoritmo 2) usa una base estándar progresiva (pasos 1 a 3), con dos pasos adicionales en cada nodo para minimizar la avaricia de los algoritmos progresivos (paso 3.3) y asegurar un *centro estructural* bien definido (3.4):

1. Se lleva a cabo una comparación por pares de todas las proteínas contra todas usando el algoritmo de MAMMOTH de pares (ver Algoritmo 3). Se obtiene una matriz de similitud NxN, donde para cada par de estructuras *i* y *j*, se calcula una puntuación de similitud como $s_{ij} = -ln(P_{ij})$, donde s_{ij} es la puntuación de MAMMOTH para el alineamiento de pares. P_{ij} es la probabilidad de que el conjunto de residuos alineados pudiera haberse obtenido por coincidencia al azar de dos plegamientos diferentes en la base de datos (Ortiz et al., 2002).
2. Se crea un dendograma aplicando un algoritmo de agrupamiento de valores promedios (*average linkage clustering algorithm*) (Johnson and Wichern, 1998) a la matriz derivada en 1.
3. Se siguen los nodos del árbol desde las hojas hasta la raíz. Así, sean A y B las dos ramas de un nodo dado, y n_A y n_B el número de estructuras que forman parte de cada rama. En cada nodo, se llevan a cabo los siguientes pasos:
   - 3.1. Se asignan correspondencias entre ambos subgrupos basándose en sus vectores promedio Cα- Cα. Se calcula la matriz de similitud, **S**, de las estructuras en A con las estructuras en B para cada par de posiciones *i* y *j* de los alineamientos parciales acumulados en cada rama, promediando las similitudes de los vectores de pares Cα-Cα sobre los n_A y n_B elementos de las ramas:

$$S_{ij}^{AB} = \frac{1}{(n_A n_B)} \sum_{k=1}^{n_A} \sum_{l=1}^{n_B} u_{ijkl} \qquad (1)$$

     donde u_{ijkl} corresponde a la similitud de la cadena principal de las posiciones *i* y *j* para las proteínas *k* y *l*, medidas por su similitud *URMS*, usando los valores previamente guardados en el paso 1. Las penalizaciones por hueco son las mismas que las utilizadas en el paso 1. El camino del alineamiento a lo largo de la matriz **S** se obtiene mediante un paso de programación dinámica local-global, para asignar correspondencias entre las estructuras n_A y n_B.
   - 3.2. Se calcula la superposición tridimensional basándose en la rutina *MaxSub* (Siew et al., 2000), (ver Algoritmo 4), implementada en MAMMOTH de pares.
   - 3.3. Se reasignan las correspondencias basándose en esta superposición tridimensional. Se calcula una nueva matriz de similitud de esta manera (Rossmann and Argos, 1976):

$$S_{ij}^{AB} = \frac{1}{(n_A n_B)} \sum_{k=1}^{n_A} \sum_{l=1}^{n_B} \left(w_b u_{ij} + w_d e^{-\alpha d_{ijkl}^2} \right) \qquad (2)$$

donde d_{ijkl} es la distancia euclídea entre los Cα de los residuos en la posiciones *i* y *j* para las proteínas *k* y *l*; α es un parámetro que controla la anchura de la gaussiana, w_b es el peso de la similitud de la cadena principal en la similitud total y w_d es el peso de la distancia de la componente de la distancia cartesiana. Una vez que la matriz se ha llenado se aplica un paso de programación dinámica para obtener las reasignaciones. Esto permite la corrección de posibles errores en el alineamiento cometidos en el paso 3.1, introduciendo información terciaria basada en la superposición inicial.

3.4. Se minimiza la fluctuación del RMSD del *centro estructural* usando una optimización SIMPLEX (Numerical-Recipies-Software, 1992). Para este último paso, se emplea un procedimiento similar al descrito por Barton y Sternberg (Barton and Sternberg, 1987). Cada una de las estructuras del subgrupo se desplaza y rota por turno, minimizando el RMSD con respecto a todos los demás miembros, que permanecen fijos. Es un proceso iterativo que se repite hasta que se alcanza la convergencia en la función de error. La función a minimizar es:

$$\varepsilon_{centro} = \sum_{m=1}^{n_{centro}} \sum_{k}^{n_A+n_B} \sum_{k\neq l}^{n_A+n_B} d\left(\vec{r}_{mk}, \vec{r}_{ml}\right)^2 \quad (3)$$

donde n_{centro} es el número de residuos en el *centro estructural*, n_A y n_B el número de elementos en los subgrupos A y B y d^2 es el cuadrado de la distancia cartesiana entre los Cα de las proteínas *k* y *l* en la posición *m* del *centro estructural* en el alineamiento. La razón de aplicar este paso es que ni en el paso 3.2 ni en el 3.3 se fuerza al *centro estructural* a estar óptimamente superpuesto, y para asegurarlo se requiere este paso adicional.

3.2.1. Algoritmo de MAMMOTH de pares

El algoritmo de alineamiento estructural múltiple MAMMOTH-mult (Lupyan et al., 2005), es una extensión de la versión de pares, MAMMOTH (Ortiz et al., 2002), desarrollada previamente en nuestro laboratorio. Por tanto, es necesario explicar los fundamentos del algoritmo original, si se quiere entender claramente el funcionamiento de la versión múltiple.

MAMMOTH es un método de alineamiento estructural rápido, totalmente independiente de secuencia, que sólo tiene en cuenta las coordenadas de los Cα y que evita todo tipo de referencia a la secuencia o a los mapas de contacto. De manera similar a otros métodos, se reduce la complejidad del problema usando una aproximación heurística: primero se encuentra el alineamiento estructural que proporciona la similitud local óptima de la cadena principal de la proteína (es decir, la similitud estructural óptima para la secuencia de aminoácidos completa para las dos proteínas que se están alineando), y luego se trata de encontrar el máximo

subconjunto de residuos por debajo de una distancia de corte predefinida en el espacio tridimensional.

```
mammoth
do i = 1 + 3 to n-3
      do j = 1 + 3 to m-3
            S(i,j) = URMS(i,j) /*Computes URMS and fills the local sim. matrix*/
      enddo
enddo
needleman-wunsch (n,m,S,M_tot,l_tot)
maxsub (n,m,M_tot,l_tot,M_core,l_core)
score = pvalue (l_core,min(n,m))
```

Algoritmo 3. Pseudocódigo de la sección principal de MAMMOTH.

```
maxsub
s_max = 0                              /*s_max is the size of the biggest subset found so far*/
for i = 1 to n-L+1
         M = {(a_i, b_i), (a_i+1, b_i+1)…(a_i+L-1, b_i+L-1)}
         M = extend(M, A, B, d)        /*Extends match size*/
         If (|M| > smax) then {s_max = |M|; M_max = M}
enddo
return M_max

extend(M, A, B, d)                     /*Extends subset iteratively*/
for j = i to k                         /*Extends  M in k = 4 iterations*/
         T = rmsfit(M)  /*Computes rot matrix of T that superimpose residues in M*/.
         N=φ
         for i = 1 to n do
            If (||a_i-T(b_i)|| < jxd/k) then
                  N = N ∪ {(a_i, b_i)}
                  M = N
            endif
         enddo
         T = rmsfit(M)
         for i = 1 to n do
            If (||a_i-T(b_i)|| > jxd/k) then
                  N = N -{(a_i, b_i)}
                  M = N
            endif
         enddo
enddo
return M
```

Algoritmo 4. Pseudocódigo de la subrutina MaxSub.

El método (ver Algoritmo 3), consiste básicamente en cuatro pasos:

1. De la traza de Cα se calcula la distancia del vector unitario de desviación cuadrática media, (*unit-vector root mean square*, URMS), entre todos los pares de heptapéptidos de ambas estructuras (Kedem et al., 1999). Esta medida es sensible a la estructura local, y fue sugerida originalmente por Chew et al. (Chew et al., 1999). Se considera una proteína descrita por su traza de Cα y para cada par sucesivo de átomos de Cα a lo largo de la cadena principal, se calcula el vector unitario en la dirección de Cα(i) a Cα(i+1). Puesto que la separación entre estos átomos consecutivos es básicamente una distancia fija (3.84 Å), esta representación captura muy bien la estructura de la cadena principal. Encadenando todos estos vectores unitarios de principio a fin se obtiene el modelo estándar de una proteína como una secuencia de Cα en el espacio. A continuación se llevan al origen de referencia todos los vectores unitarios calculados, de manera que se mapea la cadena principal en vectores en la esfera unitaria. La distancia URMS entre dos segmentos de proteína, A y B (heptapéptidos, en este caso), se puede calcular determinando la matriz de rotación que minimiza la suma del cuadrado de las distancias entre los correspondientes vectores unitarios, usando técnicas estándar (McLachlan, 1979). La raíz cuadrada de la mínima suma resultante se define como la distancia URMS entre los heptapéptidos A y B. Se ha demostrado que la métrica URMS proporciona una detección eficiente de similitudes de subestructuras en proteínas (Chew et al., 1999; Kedem et al., 1999) (Figura 4).

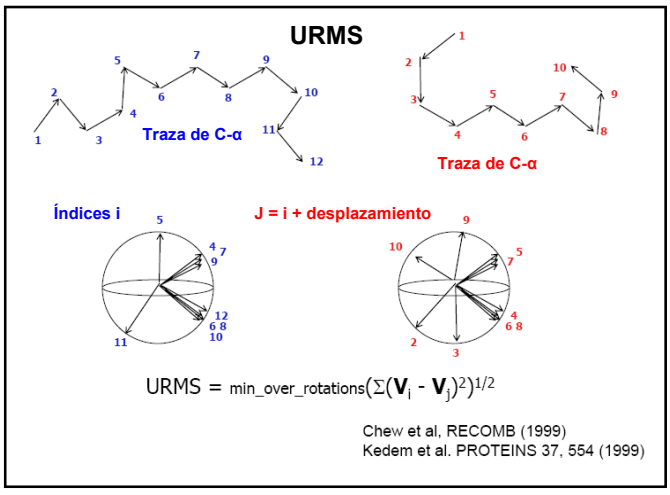


Figura 4. Esquema del cálculo del URMS.

2. Se usa la matriz derivada en el paso 1 para encontrar un alineamiento local que maximice la similitud entre las dos estructuras. Primero, se necesita transformar los valores URMS en puntuaciones de similitud. Esto se hace como en Chew et al. (Chew

et al., 1999; Kedem et al., 1999), en donde la distancia URMS mínima esperada entre dos conjuntos aleatorios de *n* vectores unitarios, ($URMS^R$), es:

$$URMS^R = \sqrt{2.0 - \frac{2.84}{\sqrt{n}}} \tag{4}$$

De la ecuación (4) se calcula la puntuación de similitud (S_{AB}) entre dos heptapéptidos, A y B, como:

$$S_{AB} = \frac{(URMS^R - URMS^{AB})}{URMS^R} \Delta(URMS^R, URMS^{AB}) \tag{5}$$

Aquí, ($URMS^R$, $URMS^{AB}$) = 10 si $URMS^R > URMS^{AB}$ y ($URMS^R$, $URMS^{AB}$) = 0 de lo contrario. Así, S_{AB} proporciona una escala de similitud entre 0 y 10. Los valores de S_{AB} se usan para construir la matriz de similitud, **S**, obtenida por comparación de todos los heptapéptidos posibles para las dos proteínas. A continuación, se aplica programación dinámica a esta matriz de similitud para construir un alineamiento de ambas estructuras basándose en similitud local de su cadena principal. Este alineamiento se produce usando un método de alineamiento global sin penalizaciones para los huecos de los extremos (*zero end gaps penalties*, (Needleman and Wunsch, 1970)). Los huecos internos se penalizan usando una función de penalización de la forma $g(k) = \alpha+\beta k$, donde *k* es el número de huecos y α y β son las penalizaciones de apertura y extensión de los mismos. Mediante ensayos de prueba y error se determinó que valores de α = 7.00 y β = 0.45 dan buenos resultados.

3. Encontrar el subconjunto máximo de estructuras locales similares, tales que sus correspondientes Cα estén cerca en el espacio cartesiano. Aquí se considera cerca, una distancia menor o igual a 4 Å. El método para encontrar este subconjunto es una pequeña variante del algoritmo heurístico de MaxSub (Siew et al., 2000) (http://www.cs.bgu.ac.il/~dfischer/MaxSub). En Maxsub (ver Algoritmo 4) dados dos conjuntos de puntos en 3D, A y B, se quiere encontrar el mayor subconjunto M, tal que para todo $(a_i, b_i) \in M$, $||a_i - T(b_i)||$ es menor que alguna distancia de corte *d*. T es la transformación que minimiza el RMSD de los residuos en M, y M puede estar formada por fragmentos no contiguos. La idea básica del algoritmo es generar "semillas" de coincidencias iniciales diferentes, de tamaño *L*. Cada una de estas coincidencias "se extiende para incluir pares adicionales. Al final se devuelve la coincidencia extendida (M_{max}) de mayor tamaño (s_{max}). Una vez que el algoritmo converge, se calcula el porcentaje de identidad estructural (*percentage of structural identity*, PSI). Éste se define como el porcentaje de residuos correspondientes a menos de 4 Å en el espacio tridimensional, medidos con respecto a la estructura más corta.

4. Se calcula la probabilidad de obtener la proporción de residuos alineados dada por azar (*P-valor*). El *P-valor* es una estimación basada en un ajuste de valor extremo de las puntuaciones resultantes de alineamientos estructurales aleatorios. Para ello, se siguió el trabajo de Abagyan y Batalov (Abagyan and Batalov, 1997).

   Se obtiene el *P-valor* como función del *z-score*:

$$P(Z > z) = 1 - e^{-e^{-(\frac{\pi}{\sqrt{6}}z+\gamma)}} \quad (6)$$

   Donde γ es la constante de Euler-Mascheroni (Gumbel, 1958).

Optimización de parámetros

Las penalizaciones para apertura y extensión de huecos para los pasos 1 y 3.1 se tomaron directamente de la versión de pares de MAMMOTH (Ortiz et al., 2002). Para la corrección de la asignación en el paso 3.3 se necesitan tres parámetros adicionales: iniciación de hueco, extensión de hueco y peso de la distancia de Cα coincidentes. Éstos, se optimizaron empíricamente con un algoritmo de templado simulado (*simulated annealing*) basado en Monte Carlo y un conjunto de datos compuesto de 105 familias usando parámetros de calidad definidos abajo, con un conjunto de alineamientos de HOMSTRAD. Se empezaron las simulaciones a partir de un conjunto de parámetros determinados por ensayo y error, y la convergencia se alcanzó después de 500 pasos de simulación. Para asignar la fiabilidad del conjunto final de parámetros se usó un test de *Jack-knife*. Los parámetros óptimos obtenidos son: *apertura de hueco* = 7.0; *extensión de hueco* = 0.15; *peso de distancia* = 8.0 y *punto de corte de distancia* = 4.0 Å. No se requirieron parámetros adicionales.

Parámetros de calidad

La calidad de los alineamientos obtenidos con MAMMOTH-mult, se midió de acuerdo a tres criterios diferentes:

1. Extensión del *centro estructural* (**% *centro***), medida como porcentaje de residuos en el *centro estructural* con respecto a la proteína más corta. Se definen dos tipos de *centro*, el "*centro estricto*" (*strict core*), y el "*centro permisivo*" (*loose core*). El ***centro estricto*** está formado por el conjunto de posiciones con un 100% de conservación (es decir, ninguna proteína presenta un hueco en esa posición) y cuya distancia Cα–Cα de unas a otras es menor de 4.0 Å en el alineamiento tridimensional final para todas las proteínas. El ***centro permisivo*** abarca aquellas posiciones con al menos el 66% de conservación y cuya distancia Cα–Cα con respecto a la media de coordenadas para esa posición en el alineamiento tridimensional es menor de 3.0 Å . Esta última definición de *centro* se usa como la función a optimizar en el paso de asignación de valor a los parámetros descrito arriba. Por otra parte, el *centro estricto* se usa en todas las comparaciones con el resto de programas (ver apartado de Resultados, pág. 67).

2. Fluctuación media del RMSD (**<*RMSD*$_{centro}$>**) de los residuos del *centro estricto*.
3. Calidad del alineamiento múltiple de secuencia correspondiente al alineamiento estructural final, calculada en términos de ***norMD*** (Thompson et al., 2001). *NorMD* es un buen parámetro de evaluación, ya que combina las ventajas de las técnicas basadas en la puntuación de las columnas del alineamiento con la sensitividad de los métodos que incorporan puntuaciones para la similitud de residuos. Además, *norMD* añade información *ab initio* de secuencia (como el número, la longitud y la similitud de las secuencias a alinear). Es importante resaltar que este parámetro sólo se usa para medir la calidad del alineamiento de salida de MAMMOTH-mult; la información de secuencia no se usa en ningún momento para construir el alineamiento estructural múltiple.

3.3. Estudio de la plasticidad del centro estructural de familias de proteínas homólogas. Análisis de componentes principales

Se usa análisis de componentes principales (*Principal Component Analysis*, PCA) (Johnson and Wichern, 1998) para extraer el conjunto de direcciones principales de movimiento que mejor describen las deformaciones experimentadas por el *centro estructural* a lo largo de la evolución. Para ello, se aplica MAMMOTH-mult a cada una de las superfamilias estudiadas, de manera que se obtiene un alineamiento estructural múltiple para cada una de ellas. De cada alineamiento se seleccionan las regiones estructuralmente conservadas para cada superfamilia (el "*centro estructural evolutivo*"). En este punto, hay dos posibilidades: considerar el *centro evolutivo* como el *centro estricto*, o bien como el *centro permisivo* de MAMMOTH-mult. Para el *centro estricto*, se obtiene una matriz $\mathbf{X}_{\mathbf{pxn}}$ que contiene las coordenadas cartesianas de los Cα que definen el centro estructural de la superfamilia, donde *n* es el número de estructuras y *p* es 3 veces el número de las posiciones de este *centro* (cada posición se define por sus correspondientes coordenadas cartesianas *x,y,z*). Si se considera el *centro permisivo*, la matriz que contiene las coordenadas de las posiciones conservadas es $\mathbf{Y}_{\mathbf{mxn}}$, donde *n* es el número de proteínas y $m = p+g$ el de posiciones del *centro permisivo* (*p* para el número de posiciones del *centro estricto* y *g* el número de posiciones de *centro estructural con huecos*). Obviamente, $m > p$ ya que el *centro permisivo* aparte de englobar al *centro estricto*, también abarca posiciones con un porcentaje de huecos dado. Al contener más posiciones, el *centro permisivo* contiene más información y puede resultar más útil que el *centro estricto*. Sin embargo, su uso presenta un problema a la hora de aplicarle la técnica estándar de PCA, en donde se necesita una matriz de partida "sin huecos"; es decir, en nuestro caso, una matriz tal que todos sus elementos estén rellenos con las coordenadas de los Cα conservados de las proteínas. En aquellas posiciones del *centro permisivo* donde alguna proteína presente un hueco, lógicamente no hay coordenadas y no es posible aplicar PCA estándar, por lo que hay que recurrir a un análisis de componentes principales mediante expectación-maximización, EM-PCA (Skocaj, 2002).

3.3.1. PCA

El objetivo principal que persigue esta técnica es la representación de las medidas numéricas de varias variables en un espacio de pocas dimensiones donde se puedan percibir relaciones que de otra manera permanecerían ocultas en dimensiones superiores. Dicha representación debe ser tal, que al desechar dimensiones superiores, la pérdida de información sea mínima.

Por tanto, el PCA reduce la dimensionalidad de los datos reteniendo tanto como sea posible la variación entre ellos. Para hacerlo, el método transforma linealmente un gran número de variables correlacionadas, a menudo conocidas como variables originales, al mismo o a un número menor de variables no correlacionadas, cada una llamada **componente principal**, de manera que las varianzas de esos componentes están en orden descendente. Así, los primeros componentes principales explican la mayor parte de la variación entre las variables originales. Seleccionados de acuerdo a un punto de corte preespecificado de porcentaje de variación total explicada, o algún otro criterio, estos componentes principales contienen toda (o la mayoría de) la información inherente en los datos y por tanto determinan la dimensionalidad de los mismos. Los componentes restantes se consideran ruido y contienen menos información. El análisis de datos se lleva a cabo basándose en estos componentes principales en lugar de en las variables originales.

En el caso que nos ocupa, de comparación del *centro estructural* de proteínas homólogas, sea $\mathbf{X_{pxn}}$ la matriz que contiene las coordenadas cartesianas de los Cα que definen el *centro estricto* de la superfamilia, donde *n* es el número de estructuras y *p* es 3 veces el número de las posiciones, la matriz de covarianzas de esta matriz es

$$\mathbf{C_{pxp}} = \frac{1}{n}\,\mathbf{X}\,\mathbf{X'} \tag{7}$$

con elementos $c_{ij} = \langle (x_i - \langle x_i \rangle)(x_j - \langle x_j \rangle) \rangle$ donde $\langle\rangle$ significa la media sobre todas las proteínas. Haciendo la descomposición espectral de $\mathbf{C_{pxp}}$ se obtiene $\mathbf{C} = \mathbf{V\Delta V'}$ donde $\mathbf{V}$ es una matriz ortogonal que contiene el conjunto de autovectores o vectores propios y $\mathbf{\Delta}$ es una matriz diagonal que contiene el conjunto de autovalores o valores propios. Si definimos $k = \min(p,n)$, se tiene $\mathbf{V_{pxk}}$ y $\mathbf{\Delta_k}$. Por tanto, la diagonal formada por los valores propios de $\mathbf{C_{pxp}}$ son $\lambda_1, \lambda_2, \ldots \lambda_k$.

Es posible reordenar de acuerdo con su magnitud los valores propios de $\mathbf{C}$ de tal manera que λ_1 sea el mayor de todos ellos, λ_2 el que le sigue, etc y λ_k el menor de todos. Esto se traduce en un reordenamiento de las columnas de la matriz $\mathbf{V_{pxk}}$ de manera que la primera sea un vector propio asociado con λ_1, la segunda un vector propio asociado con λ_2 y así sucesivamente. De esta manera, el primer vector propio $\mathbf{v_1} = (V_{11}, V_{21}, \ldots, V_{p1})$ apunta en la dirección de máxima variabilidad de la nube de puntos. Esta dirección se llama "*primera dirección principal*". El segundo vector propio $\mathbf{v_2} = (V_{12}, V_{22}, \ldots, V_{p2})$, apunta en la siguiente dirección de máxima variabilidad de la nube de puntos, llamada "*segunda dirección principal*" y así sucesivamente.

La traza de **C**, por ser suma de las varianzas de las variables originales recibe el nombre de *varianza total*, VT. Así:

$$\text{Traza}(\mathbf{C}) = \text{Traza}(\mathbf{V\Delta V'}) = \sum_{i=1}^{k} \lambda_i \tag{8}$$

Se puede probar que la varianza total es igual a la suma de los valores propios de **C** e igual a la suma de las varianzas de las componentes principales. Es decir, la varianza total es la misma con las variables originales que con las variables transformadas. Los componentes principales son vectores aleatorios no correlacionados entre sí, obtenidos mediante transformaciones lineales de las vectores de coordenadas originales.

La varianza total se descompone en un número finito de partes disjuntas λ_j de tamaños cada vez menores, lo que en la práctica proporciona un mecanismo para reducir la dimensionalidad de representación de las variables. En efecto, si se desprecian las últimas $k–r$ componentes principales, las primeras r, tendrán una tasa de representatividad igual a $(\lambda_1+ \lambda_2+_{...}+ \lambda_r)/\text{VT}\cdot 100$ % de la varianza total de las variables originales. Muchas veces, este porcentaje es bastante alto con un pequeño valor de r, lo que se traduce en una alta representatividad en un espacio de pocas dimensiones.

3.3.2. EM-PCA

Como extensión del PCA estándar, se aplicó una versión simplificada del algoritmo de análisis de componentes principales mediante expectación-maximización (*expectation-maximitation principal components analysis*, EM-PCA) (Skocaj, 2002), que permite incluir las posiciones con huecos en los alineamientos porque son tratadas como valores perdidos (*missing values*). Para esto, se aplica primero PCA estándar al alineamiento estructural de la manera usual. Así, sea $\mathbf{X_{pxn}}$ la matriz de las posiciones del *centro estricto* obtenidas del alineamiento estructural múltiple y $\mathbf{V_{pxk}}$ la matriz de autovectores y $\mathbf{\Delta_k}$ la matriz de autovalores (donde n es el número de proteínas, p es 3 veces el número de posiciones y $k = \min(p,n)$) obtenidas mediante la descomposición espectral de su matriz de covarianzas, existe una matriz de coeficientes, $\mathbf{P_{kxn}}$, tal que: $\mathbf{X_{pxn}} = \mathbf{V_{pxk}} \cdot \mathbf{P_{kxn}}$ (donde k es el número de autovectores), que transfiere las coordenadas desde el espacio PCA al espacio cartesiano original. La matriz de coeficientes, o proyección en el espacio de componentes principales, se puede escribir como:

$$\mathbf{P_{kxn}} = (\mathbf{V'_{pxk}}\,\mathbf{V_{pxk}})^{-1}\,\mathbf{V'_{pxk}}\,\mathbf{X_{pxn}} \tag{9}$$

Para construir el espacio EM-PCA, se parte de la solución de PCA, utilizando la misma matriz $\mathbf{P_{kxn}}$ de coeficientes encontrada: $\mathbf{Y_{mxn}} = \mathbf{U_{mxk}} \cdot \mathbf{P_{kxn}}$ siendo $\mathbf{Y_{mxn}}$ la matriz de posiciones del *centro permisivo*, $\mathbf{U_{mxk}}$ la nueva matriz de autovectores y m el número de posiciones del *centro permisivo*. La nueva matriz de autovectores $\mathbf{U_{mxk}}$ que definen el espacio EM-PCA, puede obtenerse como:

$$\mathbf{U_{mxk}} = \mathbf{Y_{mxn}} \cdot \mathbf{P'_{kxn}} \cdot (\mathbf{P_{kxn}} \cdot \mathbf{P'_{kxn}})^{-1} \quad (10)$$

k es el número de autovectores de PCA usados y *m* es el número de posiciones del *centro permisivo* ($m \geq p$ ya que se incluyen posiciones con huecos). En la matriz $\mathbf{Y_{mxn}}$, las posiciones correspondientes a hueco se aproximan originalmente mediante los valores promedio del resto de los elementos en la fila $<x_j>$. A diferencia de lo que se hace en las implementaciones estándar del algoritmo de EM-PCA, aquí no se itera el procedimiento para derivar una nueva matriz de coeficientes. La razón es que de esta manera, las posiciones pertenecientes al *centro estricto* en los autovectores correspondientes al *centro permisivo*, no cambian sus valores, ya que la matriz de transformación $\mathbf{P_{kxn}}$, es la misma usada en PCA y las coordenadas de las posiciones del *centro estricto* también son las mismas. De esta manera se preserva toda la información evolutiva presente en el alineamiento estructural y se incorporan aquellas posiciones que no podían usarse cuando se consideraba sólo el *centro estricto*.

3.4. Flexibilidad de proteínas. Análisis de modos normales

Se estudia si las deformaciones evolutivas observadas en las familias de proteínas están relacionadas con los modos vibracionales accesibles a su topología, para lo cual se aplica el análisis de modos normales.

Es bien conocido que las proteínas en el estado nativo, no son sistemas rígidos, sino que presentan fluctuaciones cerca de las posiciones de equilibrio. Además de estas fluctuaciones, tienen lugar otros dos tipos de transiciones conformacionales más específicas: 1) cambios a conformaciones isómeras, particularmente en cadenas laterales con enlaces rotables; y 2) cambios a gran escala: algunas proteínas pueden presentar dos o más estados de equilibrio relevantes para su función.

El análisis de modos normales (*normal mode analysis*, NMA), es una de las dos técnicas de simulación (la otra es la *dinámica molecular*, MD), usadas para estudiar la dinámica interna de las moléculas biológicas.

El análisis de modos normales se basa en un modelo armónico que asume que en el rango de las fluctuaciones térmicas, la superficie de energía conformacional puede caracterizarse por una aproximación parabólica a un único mínimo de energía. Sin embargo, existe evidencia abundante, tanto experimental (Austin et al., 1975) como computacional (Elber and Karplus, 1987), de que la aproximación armónica se rompe espectacularmente para proteínas a temperaturas fisiológicas, donde lejos de mostrar movimiento armónico en un único mínimo de energía, el sistema visita múltiples mínimos cruzando barreras energéticas de diferentes alturas. Incluso si el movimiento dentro de un único mínimo de energía fuera representativo del movimiento dentro de todos los mínimos, como parece ser el caso (Hayward et al., 1995), se

esperaría que los eventos de superación de barreras tuvieran una influencia mayor en el movimiento global de la molécula, sin ninguna relación obvia con el movimiento dentro de mínimos individuales. Dado pues, el nivel de aproximación realizado, el éxito comprobado del análisis de modos normales no deja de ser sorprendente.

Este tipo de análisis se aplicó por primera vez a las proteínas a comienzos de los años 80 usando potenciales empíricos para todos los átomos (*all-atom empirical potentials*), desarrollados para proteínas (Brooks and Karplus, 1983; Go et al., 1983). Sin embargo, el uso de este tipo de modelos y potenciales atómicos se hace impracticable cuando se aumenta el tamaño del sistema.

En los últimos años, se han desarrollado diferentes aproximaciones, siendo muy exitosos los modelos de baja granularidad (*coarse-grained models*), que simulan la dinámica vibracional de las proteínas aproximándolas a redes elásticas (Bahar et al., 1997; Tirion, 1996). En particular, en esta tesis se implementó el modelo de red anisotrópico (*Anisotropic Network Model*, ANM) (Atilgan et al., 2001).

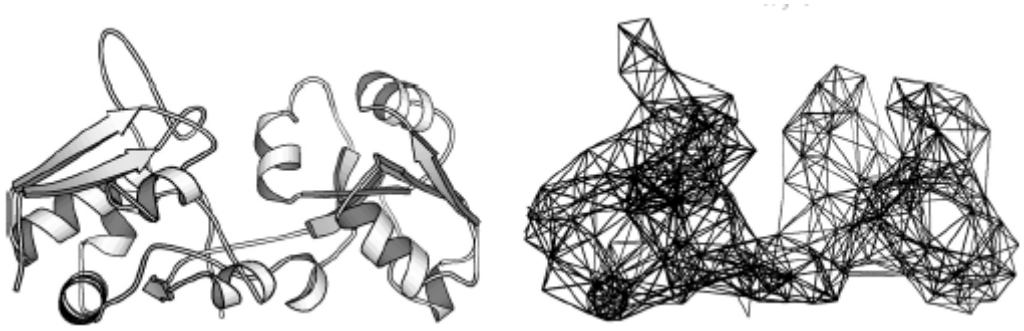


Figura 5. Representación de la estructura de la proteína de unión a lisina-arginina-ornitina (lysin-arginin-ornithine, LAO, binding protein), como una red elástica. Se muestra su estructura mediante una representación tipo "*cartoon*" (izquierda), y mediante una red de interacciones (derecha), en donde cada línea conecta pares de Cα que se encuentran a una distancia $s_{ij} <= r_c$. siendo r_c = 8 Å en este caso. (Tomado de (Tama and Sanejouand, 2001)).

3.4.1. Modelo de red anisotrópico, ANM

El ANM (Atilgan et al., 2001) es un modelo de baja resolución para estudiar la dinámica vibracional de las proteínas en el estado plegado basado en la teoría de elasticidad de redes de polímeros. Se trata de un método analítico muy eficiente cuyo postulado básico consiste en considerar la proteína en su estado plegado equivalente a una red elástica tridimensional. Los nodos de la red están sometidos a fluctuaciones gaussianas bajo el potencial de las cadenas vecinas. Los Cα equivalen a los nodos de la red y fluctúan bajo los potenciales de sus vecinos, y las interacciones entre residuos se reemplazan por muelles lineales. No se hace distinción entre diferentes tipos de aminoácidos, por lo que se adopta una constante de fuerza genérica para el potencial de interacción entre todos los pares de residuos suficientemente cerca.

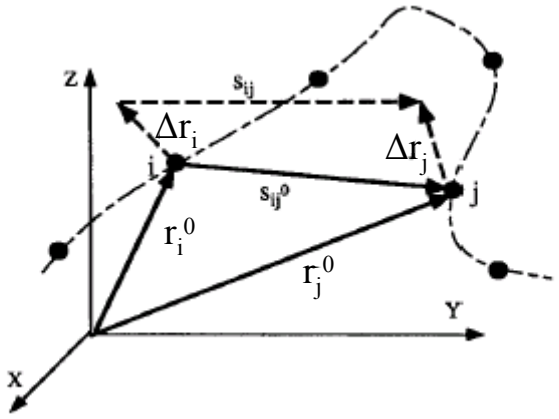


Figura 6. Representación esquemática de las fluctuaciones $\Delta\mathbf{r}_i$ y $\Delta\mathbf{r}_j$ en los vectores de posición de los residuos *i* y *j*. Los vectores de posición de equilibrio con respecto al marco *xyz* son $\mathbf{r}_i^0$ y $\mathbf{r}_j^0$, y sus valores instantáneos son $\mathbf{r}_i$ y $\mathbf{r}_j$. $\mathbf{s}_{ij}^0$ y $\mathbf{s}_{ij}$ son los vectores de separación entre los sitios *i* y *j* en el equilibrio y en el movimiento instantáneo. El cambio en la separación con respecto a las coordenadas de equilibrio es $\mathbf{s}_{ij}$ - $\mathbf{s}_{ij}^0 = \Delta\mathbf{r}_j - \Delta\mathbf{r}_i$. (Tomado de (Atilgan et al., 2001)).

El potencial conformacional total de toda la estructura se aproxima mediante un potencial armónico de la forma (Flory, 1976):

$$V = (\gamma/2)\mathbf{\Delta R}^{\mathbf{T}}\mathbf{\Gamma}\mathbf{\Delta R} \tag{11}$$

Donde $\mathbf{\Delta R}$ es la matriz que contiene los vectores de desplazamiento $\mathbf{\Delta r_i}$ de los residuos individuales del sistema y $\mathbf{\Gamma}$ es la matriz de Kirchhoff de contactos entre los residuos:

$$\Gamma_{ij} = \begin{cases} -1 & s_{ij} \le r_c \\ 0 & s_{ij} > r_c \end{cases} \tag{12}$$

donde r_c es una distancia de corte por debajo de la cual se considera que dos residuos están en contacto. Consideremos un único muelle entre dos residuos *i* y *j* sujetos al potencial armónico:

$$V = (1/2)\gamma(s_{ij} - s_{ij}^0)^2 = (1/2)\gamma([x_j - x_i)^2 + (y_j - y_i)^2 + (z_j - z_i)^2]^{1/2} - s_{ij}^0)^2 \tag{13}$$

La primera y segunda derivadas de *V* con respecto a los componentes de $\mathbf{r}_i$ son

$$\partial V / \partial x_i = -\,\partial V / \partial x_j = -\gamma(x_j - x_i)(1 - s_{ij}^0 / s_{ij}) \tag{14}$$

$$\partial^2 V / \partial x_i^{\,2} = \partial^2 V / \partial x_j^{\,2} = \gamma(1 + s_{ij}^0 (x_j - x_i)^2 / s_{ij}^3 - s_{ij}^0 / s_{ij}) \tag{15}$$

Expresiones similares se aplican a los componentes *y* y *z* de $\mathbf{r}_i$. En el equilibrio, $s_{ij} = s_{ij}^0$, y las Eqs. 14 y 15 se reducen a

$$\partial V / \partial x_i = 0 \tag{16}$$

$$\partial^2 V / \partial x_i^{\,2} = \gamma(x_j - x_i)^2 / s_{ij}^2 \tag{17}$$

Las segundas derivadas se transforman en

$$\partial^2 V / \partial x_i \partial y_j = -\partial^2 V / \partial x_j \partial y_i{}^2 = -\gamma (x_j - x_i)(y_j - y_i) / s_{ij}^2 \tag{18}$$

En el caso de Γ_{ii} los vecinos que rodean al residuo *i*, las Eqs. 17 y 18 son reemplazadas por

$$\partial^2 V / \partial x_i{}^2 = \gamma \sum_j (x_j - x_i)^2 / s_{ij}^2 \tag{19}$$

$$\partial^2 V / \partial x_i \partial y_j = \gamma \sum_j (x_j - x_i)(y_j - y_i) / s_{ij}^2 \tag{20}$$

donde las sumas se llevan a cabo sobre todos los vecinos ($j = 1, \Gamma_{ii}$) del residuo *i*.

En el caso general de *N* residuos conectados por *M* uniones, las segundas derivadas del potencial total se organizan en la matriz Hessiana $3N \times 3N$ **H**. **H** se compone de $N \times N$ super-elementos de tamaño 3×3, esto es,

$$\mathbf{H} = \begin{bmatrix} \mathrm{H}_{11} & \mathrm{H}_{12} \cdots & \mathrm{H}_{1\mathrm{N}} \\ \mathrm{H}_{21} & & \mathrm{H}_{2\mathrm{N}} \\ \vdots & & \vdots \\ \mathrm{H}_{\mathrm{N1}} & & \mathrm{H}_{\mathrm{NN}} \end{bmatrix} \tag{21}$$

El super-elemento ij^{th} ($i \neq j$) $\mathbf{H}_{\mathrm{ij}}$ de **H** es:

$$\mathbf{H_{ij}} = \begin{bmatrix} \partial^2 V / \partial x_i \partial x_j & \partial^2 V / \partial x_i \partial y_j & \partial^2 V / \partial x_i \partial z_j \\ \partial^2 V / \partial y_i \partial x_j & \partial^2 V / \partial y_i \partial y_j & \partial^2 V / \partial y_i \partial x_j \\ \partial^2 V / \partial z_i \partial x_j & \partial^2 V / \partial z_i \partial y_j & \partial^2 V / \partial x_i \partial z_j \end{bmatrix} \tag{22}$$

La Eq. 18 da los elementos de $\mathbf{H}_{\mathrm{ij}}$. Los elementos de la diagonal de super-elementos $\mathbf{H}_{\mathrm{ij}}$, vienen dados sin embargo por las Eqs. 19 (diagonal) y 20 (*off*-diagonal).

H por tanto, se puede calcular a partir de las coordenadas cartesianas de los Cα de la estructura nativa. La factorización de **H** como $\mathbf{H}=\mathbf{A}\mathbf{\Lambda}\mathbf{A}^{\mathrm{T}}$ da los *3N-6* modos normales intrínsecos (siendo *N* el número de residuos totales), correspondientes a la matriz de autovectores, **A**, cuyas frecuencias están contenidas en la matriz diagonal **Λ**.

La matriz **A** (de los modos normales), se compara con la matriz **V** (de las componentes principales del cambio evolutivo), para comprobar si existe relación entre ellas.

Por otro lado, las correlaciones cruzadas (*cross-correlations*), entre las fluctuaciones de los residuos *i* y *j* se obtienen de

$$\langle \Delta \mathbf{r}_i \cdot \Delta \mathbf{r}_j \rangle = (3k_B T / \gamma)[\mathrm{H}^{-1}]_{\mathrm{ij}} \tag{23}$$

Las fluctuaciones cuadráticas medias de los residuos individuales se pueden obtener de la Eq. 23, tomando $i = j$, esto es,

$$<\Delta\mathbf{r}_i \cdot \Delta\mathbf{r}_j> = <(\Delta\mathbf{r}_i)^2> = (3k_BT/\gamma)[\mathrm{H}^{-1}]_{ii} \quad (24)$$

$<\Delta\mathbf{r}_i \cdot \Delta\mathbf{r}_j>$ se puede expresar como la suma sobre las contribuciones $[\Delta\mathbf{r}_i \cdot \Delta\mathbf{r}_j]_k$ de los k modos individuales, en una expansión usando los autovalores λ_k y autovectores $\mathbf{a}_k$ de $\mathbf{H}$ en

$$<\Delta\mathbf{r}_i \cdot \Delta\mathbf{r}_j> = \sum_k [<\Delta\mathbf{r}_i \cdot \Delta\mathbf{r}_j>]_k = (3k_BT/\gamma) \sum_k \left[\lambda_k^{-1} a_{ij,k} a_{ij,k}^T\right] \quad (25)$$

3.5. Relación entre el espacio evolutivo y el vibracional

El estudio de la posible relación entre los modos normales de vibración obtenidos mediante ANM y las fluctuaciones estructurales evolutivas detectadas por PCA, se aborda desde dos puntos de vista: por una parte, mediante el análisis de las fluctuaciones cuadráticas medias y por otra, mediante el estudio del solapamiento entre los dos espacios.

3.5.1. Fluctuaciones cuadráticas medias

En el caso de las deformaciones evolutivas calculadas a partir de los alineamientos estructurales, la fluctuación cuadrática media para la posición k, sobre todo el conjunto de n proteínas en el alineamiento, se obtiene como:

$$\left\langle \Delta d_k^2 \right\rangle = \frac{1}{n} \sum_i^n (r_{ik} - \left\langle r_k \right\rangle)^2 \quad (26)$$

En el caso del análisis de modos normales, la fluctuación cuadrática media para cada residuo en el espacio vibracional se puede obtener a partir de la suma de los productos escalares de los *3N-6* vectores de la matriz de autovectores, escalados por el correspondiente autovalor, de la manera siguiente (Atilgan et al., 2001):

$$\left\langle \Delta d_k^2 \right\rangle = \frac{3k_BT}{\gamma} \sum_{j=1}^{3N-6} \lambda_j^{-1} \sum_{i=3k-2}^{3k} a_{ji}^2 \quad (27)$$

Se asigna un valor de 1.8 al prefactor. Y se comparan las fluctuaciones obtenidas mediante ambos métodos. Para cada superfamilia se obtuvo el coeficiente de correlación de Spearman, Rs (Langley, 1970), entre las listas de las fluctuaciones por residuo calculadas con ambos métodos. La distribución de muestreo de Rs bajo la hipótesis nula de no-correlación, se puede aproximar mediante una distribución normal teniendo E(Rs) = 0 y var(Rs)=$(n\text{-}1)^{-1}$, siendo n el número de residuos y se calculó el valor del Z-*score* de Rs como $Zscore = Rs\sqrt{n-1}$.

3.5.2. Cálculo del RMSIP

El solapamiento entre los dos espacios se calcula a partir del producto escalar de la desviación cuadrática media (*Root Mean Square Inner Product*, RMSIP) (Amadei et al., 1999) entre los vectores que componen los dos espacios y se define como:

$$\text{RMSIP} = \left(\frac{1}{D} \sum_{i=1}^{D} \sum_{j=1}^{k} (\mathbf{v}_i \cdot \mathbf{a}_j)^2 \right)^{1/2} \tag{28}$$

donde $\mathbf{v}_i$ y $\mathbf{a}_j$ son respectivamente, los autovectores del espacio evolutivo (PCA) y los del espacio vibracional (ANM); D es la dimensión del espacio evolutivo; k tiene una dimensión igual a $3N$, siendo N el número de posiciones del *centro estructural* obtenido por MAMMOTH-mult para la familia considerada.

Para simplificar las comparaciones, el espacio vibracional se restringió a sus 50 primeros modos de más baja frecuencia. De manera similar, el espacio evolutivo se restringió al número de componentes principales necesario para explicar el 70% de la varianza total observada en las estructuras, que resultó ser de unos 5 componentes en promedio para todas las superfamilias estudiadas (ver apartado de Resultados, pág. 81).

De esta manera, se obtuvo el valor del RMSIP observado entre los dos espacios. Para evaluar su significancia estadística, para cada familia se simuló una distribución empírica de valores de RMSIP bajo la hipótesis nula de no-relación entre ambos espacios (ver Figura 7).

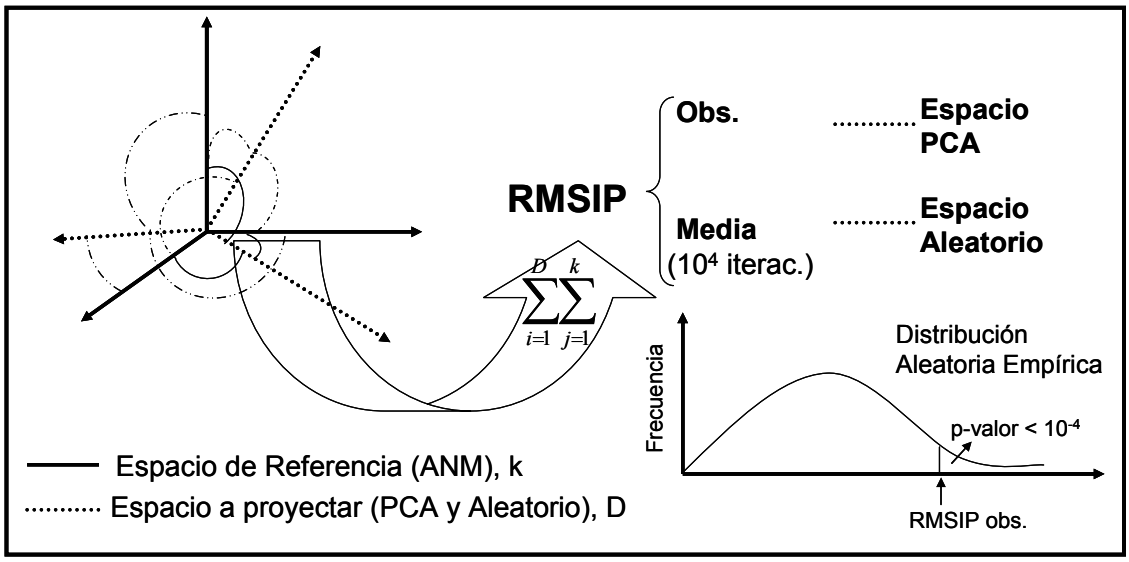


Figura 7. Esquema del cálculo del RMSIP.

Esta distribución se obtuvo proyectando sobre el espacio vibracional una serie de espacios ortogonales generados al azar, obtenidos a partir de matrices **Q** ortogonales aleatorias (*random orthogonal **Q** matrices*), siguiendo el algoritmo de Stewart (Stewart, 1980). Para construir esta distribución y asegurar su aleatoriedad, se generaron 10.000 matrices y se calculó el valor medio

del RMSIP obtenido para todas estas matrices al azar. Se determinó la significancia estadística del valor de RMSIP observado mediante el valor del *Z-score*:

$$Z - score = \frac{RMSIP(obs.) - < RMSIP(azar) >}{\sigma(azar)} \tag{29}$$

3.6. Construcción del espacio de muestreo

El espacio de muestreo es una combinación de los subespacios obtenidos mediante análisis de componentes principales (*espacio evolutivo EM-PCA*), **U**, y mediante el análisis de modos normales, (*espacio vibracional ANM*), **A**. En general, el *espacio EM-PCA* contiene la mayor parte de la información evolutiva. Sin embargo, los autovectores del *espacio de ANM* se pueden añadir al *espacio EM-PCA*, de manera que sirvan de suplemento en aquellos casos donde el muestreo estructural en el alineamiento está limitado. Para unir los dos conjuntos de autovectores, los correspondientes al *espacio ANM* se van añadiendo uno a uno al *espacio EM-PCA* y se ortogonaliza el espacio resultante cada vez, usando un procedimiento Gram-Schmidt (Arfken, 1985) (Figura 8).

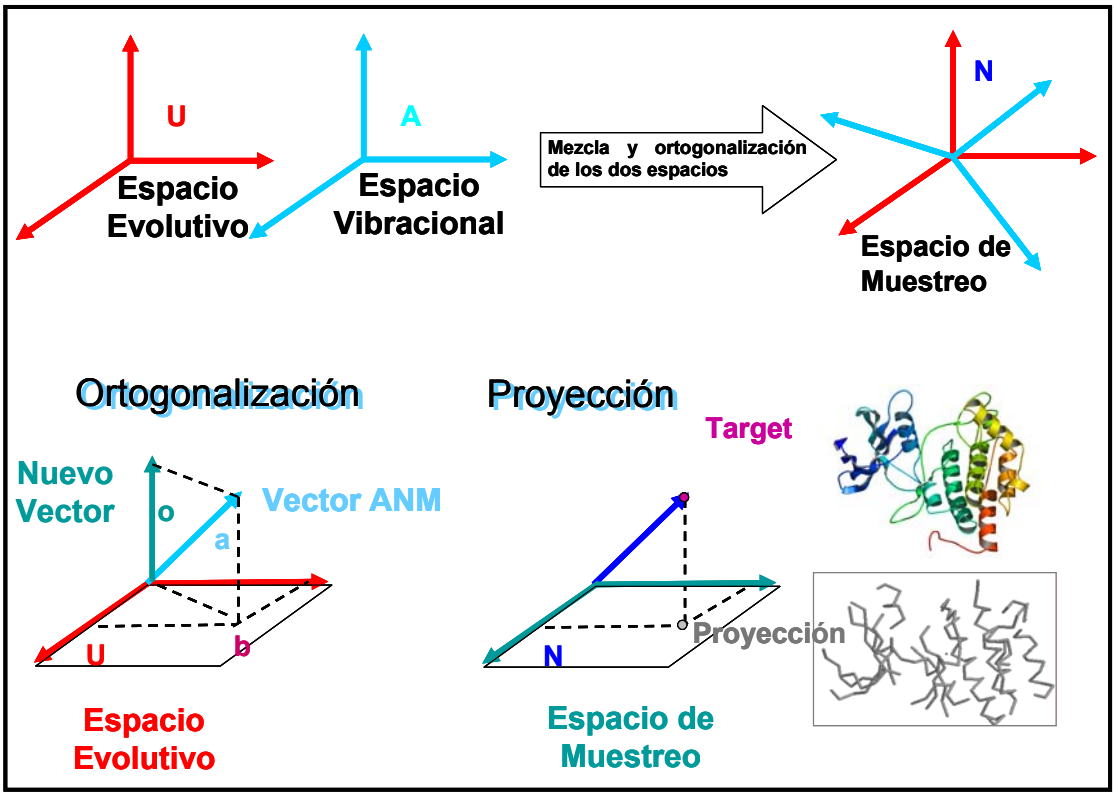


Figura 8. Esquema para la obtención del espacio de muestreo y las representaciones óptimas (proyecciones) de las proteínas problema en dicho espacio.

Dado el *espacio EM-PCA* $\mathbf{U_{mxk}}$ y uno de los vectores del *espacio ANM*, $\mathbf{a}_{m,}$ donde *m* es el número de posiciones del *centro permisivo* detectado por MAMMOTH-mult, y *k* el número de autovectores del *espacio EM-PCA*, la proyección del vector $\mathbf{a}_m$ en el espacio $\mathbf{U_{mxk}}$ se obtiene como:

$$\mathbf{b_m} = \mathbf{U_{mxk}}\,(\mathbf{U'_{mxk}}\,\mathbf{U_{mxk}})^{-1}(\mathbf{a_m}\,\mathbf{U_{mxk}}) \qquad (30)$$

y la parte ortogonal del autovector $\mathbf{a}_m$ al espacio como:

$$\mathbf{o}_m = \mathbf{a}_m - \mathbf{b_m} \qquad (31)$$

El espacio $\mathbf{U_{mxk}}$ se actualiza entonces añanadiendo $\mathbf{o}_m$ al conjunto de autovectores, dando el nuevo espacio $\mathbf{N_{mx(k+1)}}$. A continuación, se van añadiendo más vectores del *espacio ANM* al espacio **N** generado cada vez usando el mismo procedimiento. El espacio final de muestreo construido de esta manera se denomina *espacio EPA* (de *EM-PCA-ANM*). Se observa que con tan sólo 50 dimensiones, el *espacio EPA* es capaz de representar satisfactoriamente la mayoría de los centros estructurales de las proteínas estudiadas (ver apartado de Resultados, pág. 86).

3.7. Construcción de los modelos para las dianas de CASP5

Para determinar hasta qué punto este espacio EPA de 50 dimensiones podría llegar a mejorar la calidad de los modelos de proteínas en problemas reales de modelado por homología se decidió realizar una evaluación adicional de su comportamiento utilizando las mismas proteínas diana cuya estructura debían predecir los grupos participantes en CASP5. Para ello, se construyó el espacio EPA para cada una de ellas, se proyectó su estructura nativa en él, se reconstruyó el modelo completo a partir de esta proyección y se evaluó su calidad final (ver Figura 9).

3.7.1. Proyección de la diana

Una vez que se ha definido el espacio de muestreo, EPA, para cada diana se determina analíticamente la traza de Cα de su *centro estructural* en este espacio, mediante un ajuste por mínimos cuadrados estándar.

3.7.2. Reconstrucción del modelo completo a partir de la proyección

A partir de la proyección obtenida en el paso anterior, se construye un modelo lo más completo posible para cada diana. La cadena principal y las cadenas laterales se construyeron usando el paquete MMTSB (Feig, 2001) y SCWRL (Canutescu et al., 2003), respectivamente, mientras que los *lazos* de menos de 6 residuos se añadieron con MODELLER (Fiser et al.,

2002). Finalmente, se llevó a cabo una minimización de la energía con AMBER8 (D.A. Case, 2005) para evitar los posibles choques existentes entre átomos.

3.7.3. Evaluación del modelo reconstruido a partir de la proyección

Para evaluar la calidad de los modelos obtenidos para las dianas de CASP5, además de evaluar sus RMSD´s con respecto a sus estructuras experimentales, se asignó la geometría de los modelos con PROCHECK (Laskowski et al., 1993) y se calculó el porcentaje de estructura modelada, porcentajes de ángulos chi-1 y chi-2 correctos, posición de los ángulos torsionales de la cadena principal en el mapa de Ramachandran y número de malos contactos entre átomos (Figura 9).

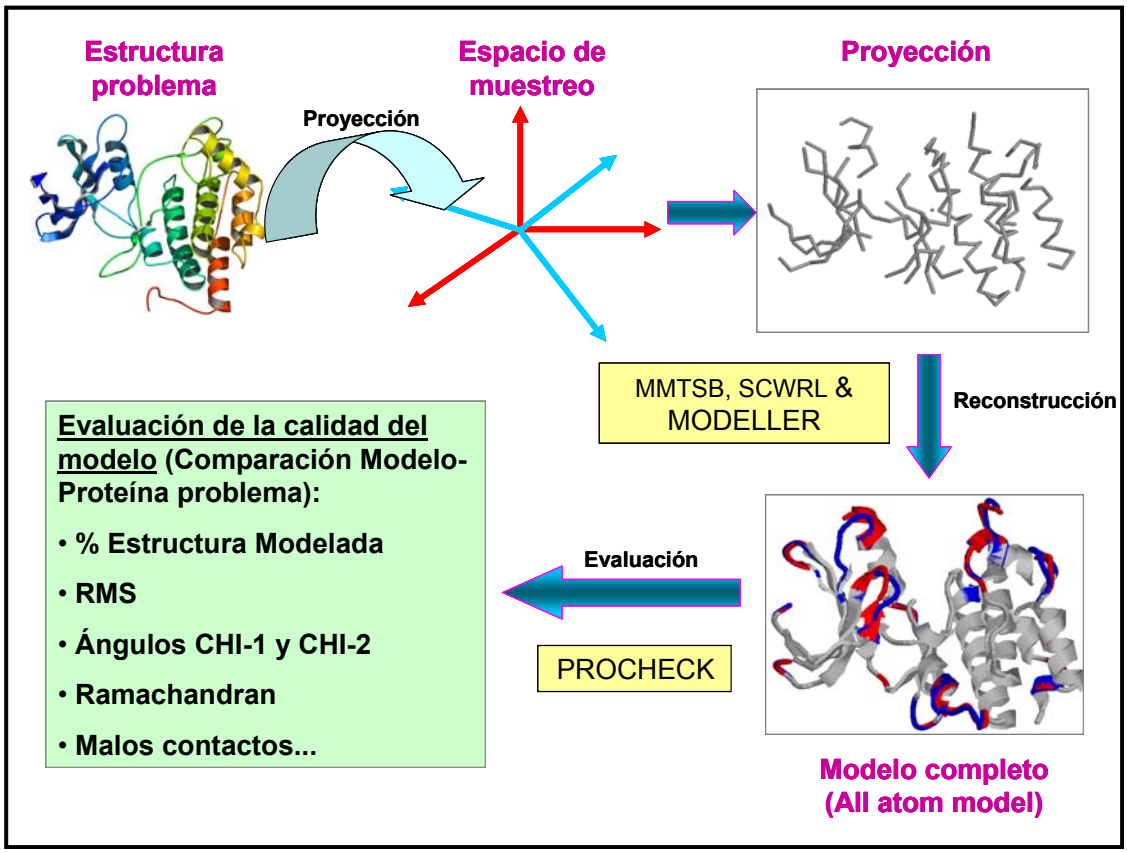


Figura 9. Esquema del protocolo general de evaluación de los modelos generados.

De manera adicional, se evaluó la calidad del espacio EPA para representar las dianas de CASP5, comparando los modelos obtenidos, con los que se obtuvieron usando la aproximación FRAGBENCH (Contreras-Moreira et al., 2005) para cada diana. Se realizó la comparación en términos del parámetro GDT_TS. En la aproximación de FRAGBENCH, se identificaron los mejores modelos basados en fragmentos para cada una de las dianas de una colección de patrones estructuralmente similares disponible en el momento en que CASP5 tuvo lugar. El

único cambio en el protocolo con respecto al artículo original (Contreras-Moreira et al., 2005), fue el uso de la última versión del programa LGA (03/2005, (Zemla, 2003)). GDT_TS es el principal parámetro de evaluación usado en los experimentos de CASP. La puntuación GDT_TS mide la similitud entre dos estructuras basándose en una combinación de las fracciones de residuos que después de la superposición de las dos estructuras a comparar, están dentro de unas distancias de corte de 1, 2, 4 y 8 Å. Se calcula como (*P1*+*P2*+*P4*+*P8*)/4, donde *Pn* es el porcentaje de residuos en el modelo que están a una distancia menor de *n* Å de los correspondientes residuos en la diana después de la superposición.

3.8. Generación de conformaciones. Simulaciones de intercambio de réplicas de Monte-Carlo

Una vez definido el espacio de muestreo, EPA y comprobada su calidad para representar satisfactoriamente las estructuras de las proteínas, se realiza una búsqueda conformacional en él para localizar conformaciones que se encuentren en los mínimos de la superficie de energía asociada.

En este apartado se explican las estrategias utilizadas normalmente en la búsqueda conformacional y la forma particular en la que hemos abordado este problema en esta tesis mediante el uso combinado del espacio EPA de muestreo y las simulaciones de intercambio de réplicas de Monte Carlo como algoritmo de búsqueda en él.

3.8.1. Búsqueda conformacional

Para llevar a cabo la búsqueda conformacional, si fuera posible, lo ideal sería identificar todos los mínimos de energía presentes en el espacio; sin embargo, en el caso de las proteínas, el número de mínimos es tan elevado que resulta imposible encontrarlos todos. Para tratar de hacerlo, se han desarrollado varios métodos que utilizan estrategias diferentes: desde aquellos que realizan una búsqueda exhaustiva y sistemática por todo el espacio, hasta aquellos que emplean aproximaciones aleatorias. Al contrario que en la búsqueda sistemática, donde se explora la superficie de energía de una molécula de una manera predecible, en las búsquedas aleatorias no es posible predecir el orden en el que se van a generar las conformaciones. Una búsqueda aleatoria se puede mover desde una región de la superficie de energía a otra completamente desconectada de la primera en un solo paso. Usualmente, los métodos de búsqueda aleatoria, se conocen como métodos de simulación de Monte Carlo (Kalos, 1986) para resaltar su base estocástica, es decir, no-determinista, en contraposición a otros métodos de simulación deterministas (como por ejemplo, la dinámica molecular). Y fueron llamados así por su naturaleza aleatoria, en honor del famoso casino del principado de Mónaco, considerado como "*la capital del juego de azar*". Una simulación de Monte Carlo genera configuraciones de

un sistema cambiando al azar las posiciones de las especies presentes, conjuntamente con sus orientaciones y conformaciones.

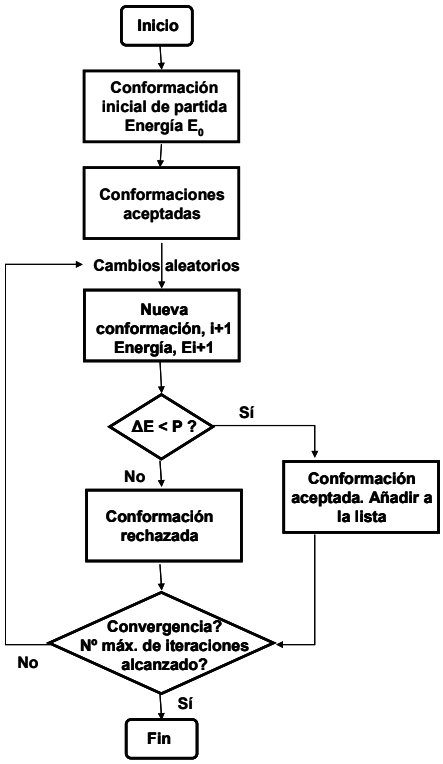


Figura 10. Esquema general de las etapas en la búsqueda conformacional utilizando Metropolis Monte-Carlo. ΔE = $\exp\left[-(E\ (r_{nueva}{}^{N}) - E\ (r_{antigua}{}^{N}))/k_B T\right]$; P es un número aleatorio entre 0 y 1.

En cada iteración, se hace un cambio aleatorio a la conformación actual y se evalúa su energía, comparándola con un valor de referencia; si se acepta, la nueva conformación se guarda y se selecciona una de las conformaciones guardadas como punto de partida para la siguiente iteración del algoritmo. El proceso continúa hasta que se haya alcanzado un número máximo de iteraciones o hasta que se haya cumplido un criterio de convergencia dado. El **algoritmo de Metropolis Monte Carlo** (Metropolis, 1953), se usa a menudo para hacer la selección de conformaciones. Cada nueva estructura generada se acepta como punto de partida para la siguiente iteración si su energía es menor que la de la estructura previamente aceptada o si el factor de Boltzmann de la diferencia de energía $\exp\left[-(E\ (r_{nueva}{}^{N}) - E\ (r_{antigua}{}^{N}))/k_B T\right]$, es mayor que un número aleatorio entre 0 y 1. Los pasos del algoritmo se esquematizan en el diagrama de la Figura 10.

En este método se comunican movimientos relativamente largos al sistema y se determina si la estructura resultante alterada es energéticamente posible a la temperatura de simulación. De esta manera, el sistema salta abruptamente de conformación en conformación, en vez de evolucionar de forma suave a lo largo del tiempo y puede atravesar barreras sin problema, lo único que importa es la energía relativa de las conformaciones antes y después del salto.

Método de Metrópolis Monte-Carlo con templado simulado

Una técnica usualmente empleada junto con las búsquedas de Metropolis Monte Carlo es el llamado *templado simulado* (*simulated annealing*) (Kirkpatrick, 1983) (ver Figura 11). El *templado* es un proceso en el cual la temperatura de una sustancia en estado fundido se reduce lentamente hasta que el material cristaliza. Es muy importante hacer un uso cuidadoso del control de temperatura para llegar a obtener el cristal perfecto, el cual se corresponde con el mínimo global de la energía libre del sistema. El *templado simulado* es un método computacional que imita este proceso para encontrar la solución óptima o mejor, a problemas que pueden presentar un gran número de soluciones posibles.

En este método, una función de puntuación desempeña el papel que tiene la energía libre en un templado físico, y un parámetro de control el de la temperatura. A altas temperaturas, el sistema es capaz de ocupar regiones de alta energía del espacio conformacional y superar barreras energéticas muy altas. A medida que la temperatura decae, los estados de menor energía se hacen más probables, de acuerdo a la distribución de Boltzmann. En el cero absoluto, el sistema debería ocupar el estado de menor energía (es decir, la conformación del mínimo global). Para garantizar que se alcanza la solución óptima, se requeriría un número infinito de pasos de temperatura, en cada uno de los cuales se debería dejar al sistema alcanzar el equilibrio térmico; además, es necesario también un cuidadoso control de la temperatura cuando la energía del sistema es comparable con la altura de las barreras que separan una región del espacio conformacional de otra.

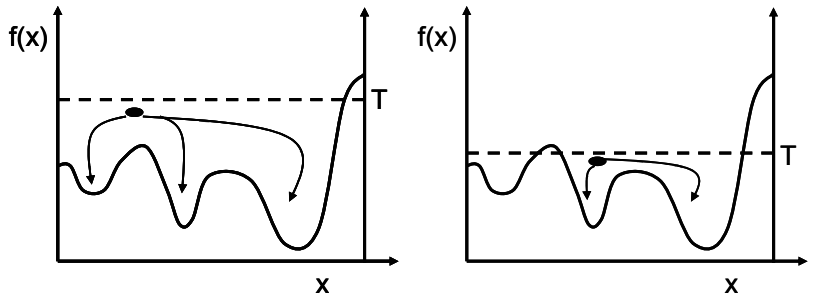


Figura 11. Algoritmo de Metropolis aplicado en un método de templado simulado. A medida que la temperatura (T) del sistema decae, el número de mínimos locales disponibles para evaluar la función f(x) disminuye. A la temperatura inicial, hay tres mínimos accesibles (izquierda), mientras que a una temperatura menor, sólo son accesibles dos (derecha).

Si existen barreras de energía muy elevada entre los mínimos de la superficie potencial, puede ocurrir que una simulación de Monte Carlo con templado simulado convencional se quede atrapada en unos pocos de estos mínimos y no muestree adecuadamente regiones más amplias del espacio térmicamente accesible. Por lo que los resultados dependerán fuertemente de las condiciones iniciales. Esta simulación puede aparentar poseer todas las carácterísticas de

una buena simulación en términos de su convergencia, pero puede dar resultados completamente incorrectos. Se dice entonces que esta simulación es *quasi-ergódica*. A bajas temperaturas, la simulación es incapaz de cruzar las barreras de alta energía debido al factor de Boltzmann favorable. Para intentar superar este problema, se pueden emplear diferentes métodos, entre ellos el de intercambio de réplicas (*Replica Exchange Monte-Carlo*, REMC) (Hukushima and Nemoto, 1996; Hukushima et al., 1996), que fue el que se implementó en esta tesis.

Método de intercambio de réplicas de Monte-Carlo

El método de intercambio de réplicas de Monte Carlo (REMC) (Hukushima and Nemoto, 1996; Hukushima et al., 1996), o templado paralelo (*parallel tempering*), se implementó en esta tesis para realizar la búsqueda conformacional. REMC es una herramienta poderosa para muestrear espacios multidimensionales, pero sólo se había empleado hasta el momento en la simulación de péptidos, rara vez en la simulación de proteínas, debido a su elevado número de grados de libertad, ya que el número de simulaciones (temperaturas) paralelas necesarias con este método crece rápidamente con él, haciendo imposible la aplicación de REMC. No obstante, puesto que el espacio EPA reduce considerablemente la dimensionalidad del muestreo, el número de grados de libertad del sistema disminuye y con él, el número de simulaciones paralelas necesarias, por lo que esta técnica resultó aplicable en los muestreos realizados en este trabajo.

En el método REMC estándar, se tienen M réplicas del sistema a diferentes temperaturas T_m (m = 1, …., M), y se lleva a cabo en dos fases que se iteran: primero, cada réplica muestrea durante un cierto número de pasos, de forma independiente a su temperatura correspondiente; después, al final de esta fase, se intercambian las réplicas que están a temperaturas vecinas de acuerdo a la probabilidad de su distribución. Sea $X = \left\{x_m^{[i]}, x_n^{[j]}\right\}$ el estado de la replica i en la temperatura T_m y la réplica j a la temperatura T_n. La probabilidad de intercambio entre $X = \left\{x_m^{[i]}, x_n^{[j]}\right\}$ y $X' = \left\{x_m^{[j]}, x_n^{[i]}\right\}$ viene dada por

$$w(X \to X') = \begin{cases} 1, & para \Delta \le 0 \\ \exp(-\Delta), & para \Delta > 0, \end{cases} \tag{32}$$

donde $\Delta = (\frac{1}{T_n} - \frac{1}{T_m})(E_i - E_j)$, E_i y E_j son las energías para las réplicas i y j de acuerdo a una función de energía dada. De esta manera, las estructuras pueden superar las barreras de energía potencial a altas temperaturas, y alcanzar los mínimos locales/globales a las bajas. En esta tesis, las temperaturas se distribuyeron siguiendo el método de Okamoto (Mitsutake et al., 2003), como $T_i = T_1(\frac{T_N}{T_1})^{(\frac{i-1}{N-1})}$, donde T_1 y T_N son las temperaturas más alta y más baja,

respectivamente, y N es el número de temperaturas o réplicas. En la implementación realizada, $T_1 = 2.0$ y $T_N = 0.05$; $N = 8$ y el número de pasos de simulación fue de 10^5.

Las regiones de *centro estructural* y de *lazos* se simularon por separado. En primer lugar, se optimiza la región correspondiente al *centro*, realizando movimientos en el *espacio EPA* y a continuación, se optimizan los *lazos* en un *espacio no-EPA*, dejando fija la estructura del *centro* optimizada anteriormente. Los detalles de estas simulaciones se describen en los siguientes apartados.

3.8.2. Simulación del centro estructural

Sea $\mathbf{N_{mxk}}$ el *espacio EPA* de muestreo (donde k es el número de vectores que lo forman y m el número de posiciones del *centro permisivo* determinado por MAMMOTH-mult). Se sitúa su origen de coordenadas es la estructura promedio de todas las proteínas de la familia. Cualquier conformación puede representarse en este espacio y relacionarse con el espacio 3D según:

$$\mathbf{Y_{mxn}} = < \sum_n \mathbf{Y_{mxn}} > + \mathbf{N_{mxk}} \cdot \mathbf{P_{kxn}} \tag{33}$$

Donde:

m = Nº de posiciones del *centro estructural*

n = Nº total de estructuras

k = Nº de autovectores

$\mathbf{Y_{mxn}}$ = Coordenadas de la diana en el espacio 3D

$< \sum_n \mathbf{Y_{mxn}} >$ = Coordenadas de la estructura promedio en el espacio 3D

$\mathbf{P_{kxn}}$ = Coordenadas de la diana en el *espacio EPA*. Son los coeficientes de ajuste de mínimos cuadrados al realizar la proyección.

$\mathbf{N_{mxk}}$ = Matriz de vectores del espacio EPA.

Para realizar la simulación sin salirse del *espacio EPA*, hay que cambiar únicamente la matriz de coeficientes, **P**. Si se cambiase directamente la matriz **Y** se estarían realizando modificaciones en el espacio 3D y por tanto, los movimientos se estarían realizando fuera del *espacio EPA*. La magnitud de los movimientos en el *espacio EPA* se puede relacionar con la de los movimientos en el *espacio 3D* de la siguiente manera:

Sea $RMSD_{ij}$ la desviación cuadrática media entre las estructuras i y j en el espacio 3D:

$$RMSD_{ij}^2 = \frac{1}{(m/3)} \sum_{l=1}^{m} \left[Y_{lxi} - Y_{lxj} \right]^2 \tag{34}$$

Puesto que el *espacio EPA* es ortonormal, el *RMSD* en función de las coordenadas del mismo queda como:

$$RMSD_{ij}^2 = \frac{1}{(m/3)} \sum_{l=1}^{k} \cdot \left[P_{lxi} - P_{lxj} \right]^2 \tag{35}$$

Por tanto, mover la estructura 1 unidad en el *espacio EPA* es:

$$1 = \sqrt{\sum_{l=1}^{k} \left[P_{lxi} - P_{lxj} \right]^2} \tag{36}$$

lo que equivale en el espacio 3D a un *RMSD* de:

$$RMSD_{ij} = \frac{1}{\sqrt{(m/3)}} \tag{37}$$

Los límites del movimiento en los autovectores del *espacio EPA* se determinaron a partir de 3 veces la máxima desviación de las estructuras desde la promedio. Para ello, se proyecta cada estructura del conjunto a lo largo de cada dirección, **k**, se toman los coeficientes de esta proyección y si d_k es el mayor de ellos en valor absoluto, se establece [$-3d_k$, $3d_k$] como los límites del movimiento en el autovector **k**. Se establece también un límite de escalado máximo de 5 Å para todos los autovectores, dado que las estructuras en todos los miembros de la familia tienen un *RMSD* menor de este valor con respecto a la estructura promedio, lo que equivale a limitar el muestreo en el rango $\left[-5\sqrt{(m/3)}, 5\sqrt{(m/3)} \right]$ para cada autovector.

La modificación de las coordenadas sujetas a deformación se obtiene como:

$$\mathbf{Y}_{\mathbf{mxn}}^{\mathbf{Nuevo}} = \mathbf{Y}_{\mathbf{mxn}} + \sum_{l=1}^{k} \mathbf{N}_{\mathbf{mxl}} \cdot \mathbf{D}_{\mathbf{l}}$$

donde **D** contiene los coeficientes de deformación de la estructura en cada autovector. Éstos se calculan como:

$$\mathbf{D_k} = n\_aleatorio \cdot amplitud \tag{39}$$

siendo:

n_aleatorio =>Número aleatorio

amplitud => Constante

Los mejores resultados se obtuvieron para *n_aleatorio* $\in$ (-1, 1) y *amplitud* $\in$ (0.1, 1.0).

La nueva conformación se acepta o rechaza según la función de energía adoptada (ver apartado 3.8.5) y a continuación tiene lugar el siguiente paso de la simulación REMC. Las nuevas conformaciones se van generando mediante esta actualización de coordenadas partiendo cada vez de la anterior conformación previamente aceptada.

3.8.3. Simulación de lazos

Modelar un *lazo* requiere satisfacer la limitación de conectar los dos segmentos de proteína en los extremos con una conformación peptídica físicamente razonable. Los residuos fijos en los extremos izquierdo y derecho del *lazo* a modelar son los residuos de anclaje N- y C-terminales, respectivamente. Estos puntos de anclaje suponen una limitación en cuanto a las conformaciones posibles del *lazo*, ya que reducen el tamaño del espacio conformacional accesible, pero satisfacer esta limitación representa un reto algorítmico considerable. Este problema, introducido por primera vez por Go y Scheraga (Go, 1970), es conocido como el "*problema del cierre de lazos*" (*loop closure problem*) y ha sido objeto de una investigación intensiva a lo largo de los años debido a su gran importancia en predicción estructural.

En el caso de *lazos* de tres residuos, el problema puede resolverse analíticamente mediante diferentes métodos (Bruccoleri, 1985; Go, 1970; Wedemeyer, 1999). Otra aproximación consiste en usar librerías de fragmentos derivadas de un conjunto de estructuras de proteínas resueltas y buscar fragmentos o combinaciones de éstos susceptibles de unir los dos extremos fijos (Jones and Thirup, 1986; Rohl et al., 2004). Más recientemente, se ha abordado este problema mediante el uso de algoritmos tomados del campo de la robótica, en particular de la cinemática inversa (Coutsias, 2004; Kolodny, 2005; Manocha, 1994), y se ha demostrado que proporcionan buenos resultados cuando se aplican al caso de *lazos* más grandes, típicamente de 4 a 15 residuos (Wang, 1991).

En cinemática, una proteína se puede considerar como una cadena de *K* uniones y *K+1* eslabones rígidos. La mayor parte de las veces, las longitudes y ángulos de enlace se consideran constantes, mientras que los ángulos dihedros (Φ y Ψ), se pueden cambiar (Figura 12).

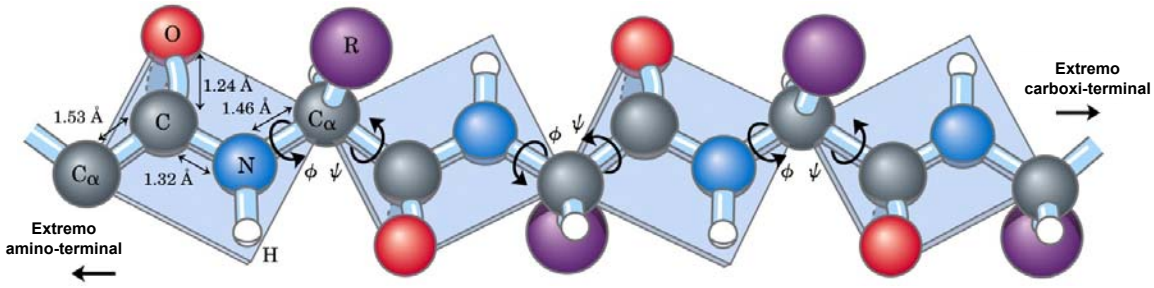


Figura 12. Grados de libertad torsionales (Φ y Ψ), a lo largo de una cadena de residuos (tomado de (Alberts, 1994)).

En los problemas de cinemática directa, dados los ángulos entre los eslabones de la cadena, se pretende encontrar la posición del final de la estructura. La cinemática inversa es justo lo contrario a esto: dada una cadena de uniones en serie y la posición relativa de un extremo con respecto al otro, se pretenden encontrar todos los posibles valores de los parámetros de unión entre los dos extremos. Es lo que se conoce como el "*problema de la*

terminación" (*the completion problem*). En el contexto de la determinación estructural, los datos de entrada para resolver este problema son una estructura parcial de proteína, dos residuos de anclaje y la secuencia de aminoácidos del fragmento a modelar. La salida, debería consistir en unas pocas conformaciones candidatas del fragmento, que respeten la limitación del cierre (asegurando que el principio y el final del fragmento coinciden exactamente con los residuos de anclaje), que eviten choques entre átomos y que satisfagan otras limitaciones (como por ejemplo, minimizar una función de energía (T en la Figura 13), que mida la calidad de la conformación).

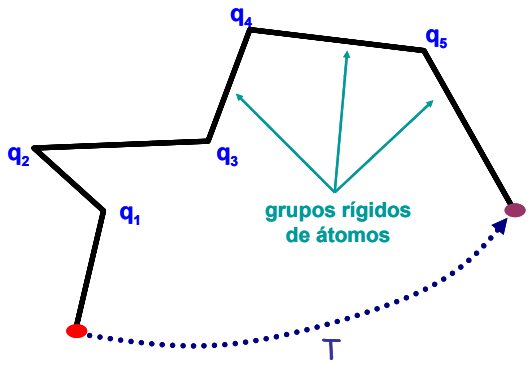


Figura 13. Representación esquemática de un *lazo* con sus extremos N- y C- terminales resaltados en rojo y magenta, respectivamente.

Existe un algoritmo lo suficientemente flexible, fácil de programar, conceptualmente simple y rápido, que da muy buenos resultados para resolver el problema del cierre de *lazos* y es el que se utilizó en esta tesis. Se trata del algoritmo CCD (*cyclic coordinate descent*) (Canutescu and Dunbrack, 2003).

Algoritmo CCD

En la implementación del algoritmo CCD (Canutescu and Dunbrack, 2003) realizada en esta tesis, los puntos de anclaje N- y C- coinciden con los residuos N y C- terminales del *lazo* y permanecen fijos a lo largo del cálculo (Figura 14).

Con la ayuda de los Cα de la cadena principal como puntos de anclaje, se construyen los *lazos* a partir de la región conservada del *centro estructural*. Las longitudes y ángulos de enlace usadas se obtuvieron de AMBER8 (D.A. Case, 2005).

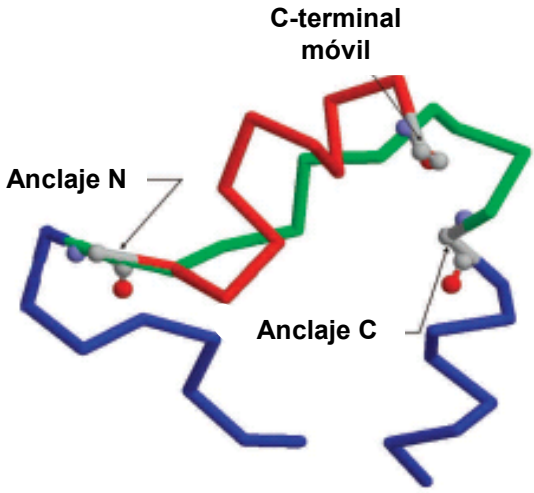


Figura 14. Traza de Cα del *lazo* antes (rojo), y después (verde), del cierre, con las estructuras secundarias flaqueándolo en azul. Se indican los anclajes N y C fijos, y el extremo C-terminal móvil. El problema del cierre del *lazo* consiste en ajustar los grados de libertad de los ángulos torsionales, de manera que el extremo móvil C-terminal se superponga con el punto de anclaje fijo C-terminal (tomado de (Canutescu and Dunbrack, 2003)).

El método consiste en ajustar un ángulo torsional cada vez para minimizar la distancia entre la posición actual del C-terminal y su posición deseada (la del anclaje C). Así, en cada paso del método CCD, el problema de minimización original *n*-dimensional se reduce a un simple problema unidimensional (Figura 15). El algoritmo va iterativamente a través de todos los ángulos torsionales ajustables desde el extremo C- hasta el extremo N-terminal del *lazo*.

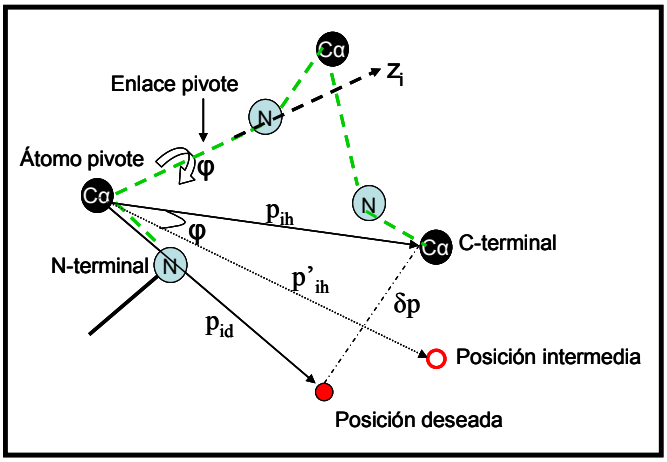


Figura 15. Un paso del algoritmo CCD. Adaptado de (Sharma et al., 2005).

En cada paso del algoritmo CCD el enlace alrededor del cual se efectúa la rotación se llama "*enlace pivote*", y su átomo precedente es el "*átomo pivote*". El ángulo torsional correspondiente al *enlace pivote* es el que se tiene que determinar.

Sea $\mathbf{p}_{id}$ el vector de posición entre el *átomo pivote* y la posición deseada del C-terminal; $\mathbf{p}'_{ih}$ es el vector de posición entre el *átomo pivote* y un estado intermedio alcanzado por el C-terminal durante la transición; $\mathbf{p}_{ih}$ es el vector de posición entre el *átomo pivote* y la posición actual del C-terminal. Entonces **δp** denota el vector de error entre las posiciones deseada y actual del C-terminal y φ es el ángulo que se rota el *enlace pivote*.

Los errores entre las localizaciones actual y deseada del C-terminal, se pueden describir mediante la siguiente función escalar positiva de φ que representa el error escalar de posición (distancia) Δp, que se usa como función objetivo a minimizar:

$$\Delta p(\varphi) = \boldsymbol{\delta p}(\varphi)\cdot \boldsymbol{\delta p}(\varphi) = (\mathbf{p}_{id} - \mathbf{p}'_{ih}(\varphi))\cdot(\mathbf{p}_{id} - \mathbf{p}'_{ih}(\varphi)) \quad (40)$$

En la ecuación de arriba, el símbolo "·" representa el producto escalar de dos vectores.

El vector $\mathbf{p}'_{ih}$ se obtiene rotando $\mathbf{p}_{ih}$ un ángulo φ alrededor del *enlace pivote*:

$$\mathbf{p}'_{ih}(\varphi) = [R(\mathbf{z}_i,\varphi)]\ \mathbf{p}_{ih} \quad (41)$$

donde $[R(\mathbf{z}_i,\varphi)]\ \mathbf{p}_{ih}$ es una matriz de rotación 3x3 y $\mathbf{z}_i$ es el vector unitario a lo largo de $\mathbf{z}_i$. Expandiendo la Ec. 40 usando la Ec.41, se obtiene la siguiente ecuación:

$$\Delta p(\varphi) = \mathbf{p}_{id}\cdot \mathbf{p}_{id} + \mathbf{p}_{ih}\cdot\mathbf{p}_{ih} - 2\mathbf{p}_{id}\cdot([R(\mathbf{z}_i,\varphi)]\ \mathbf{p}_{ih}) \quad (42)$$

Puesto que los productos $\mathbf{p}_{id}\cdot \mathbf{p}_{id}$ y $\mathbf{p}_{ih}\cdot\mathbf{p}_{ih}$ son constantes positivas, minimizar Δp , es lo mismo que maximizar:

$$g(\varphi) = \mathbf{p}_{id}\cdot([R(\mathbf{z}_i,\varphi)]\ \mathbf{p}_{ih}) \quad (43)$$

La parte derecha de la Eq. 43 se puede expandir como:

$$[R(\mathbf{z}_i,\varphi)]\ \mathbf{p}_{ih} = \mathbf{z}_i(\mathbf{p}_{ih}\cdot\mathbf{z}_i)(1-\cos(\varphi)) + \mathbf{p}_{ih}\cos(\varphi) + (\mathbf{z}_i \times \mathbf{p}_{ih})\mathrm{sen}(\varphi) \quad (44)$$

Sustituyendo la Eq. 44 en la Eq. 43, $g(\varphi)$ se transforma en:

$$g(\varphi) = (\mathbf{p}_{id}\cdot\mathbf{z}_i)(\mathbf{p}_{ih}\cdot\mathbf{z}_i)(1-\cos(\varphi)) + (\mathbf{p}_{id}\cdot\mathbf{p}_{ih})\cos(\varphi) + \mathbf{p}_{id}\cdot(\mathbf{z}_i \times \mathbf{p}_{ih})\mathrm{sen}(\varphi) \quad (45)$$

Y de aquí:

$$g(\varphi) = k_1(1-\cos(\varphi)) + k_2\cos(\varphi) + k_3\,\mathrm{sen}(\varphi) \quad (46)$$

Donde k_1, k_2 y k_3 son coeficientes constantes dados por:

$$k_1 = (\mathbf{p}_{id}\cdot\mathbf{z}_i)(\mathbf{p}_{ih}\cdot\mathbf{z}_i);\ k_2 = (\mathbf{p}_{id}\cdot\mathbf{p}_{ih}) \text{ y } k_3 = \mathbf{p}_{id}\cdot(\mathbf{z}_i \times \mathbf{p}_{ih}) \quad (47)$$

Si no hay limitaciones de entorno impuestas en φ, entonces $g(\varphi)$ se maximiza si se satisfacen las siguientes condiciones:

$$\frac{dg(\varphi)}{d(\varphi)} = (k_1 - k_2)sen(\varphi) + k_3\cos(\varphi) = 0 \quad (48)$$

$$\frac{d^2 g(\varphi)}{d\varphi^2} = (k_1 - k_2)\cos(\varphi) - k_3 sen(\varphi) < 0 \quad (49)$$

De estas ecuaciones se determina un único valor de φ. De la Ec. 48 se puede calcular el ángulo φ:

$$\varphi = \tan^{-1}\left[\frac{k_3}{k_2 - k_1}\right] \tag{50}$$

Esto determina un valor posible para φ_c en el rango $-\pi/2 < \varphi_c < \pi/2$. Sin embargo, puesto que la función *tan* es periódica, hay además otro valor potencial a considerar: $\varphi_c + \pi$. De estos valores candidatos, aquellos que pasen la prueba de la segunda derivada en la inecuación 49 son los valores maximizadores de la función objetivo de la Ec. 40. Si hay más de un valor, la función objetivo se evalúa con cada uno de ellos para determinar cuál da el máximo valor verdadero.

Una vez que φ se ha determinado inequívocamente de esta manera, se rota el correspondiente *enlace pivote* para reflejar el cambio y se calcula la posición del C-terminal usando técnicas de cinemática directa. A continuación, se procede a una nueva iteración del método CCD.

3.8.4. Resumen de los tipos de movimientos realizados en la simulación

Los tipos de movimientos realizados en la simulación REMC de las estructuras para la generación de los modelos finales se resumen en la Figura 16.

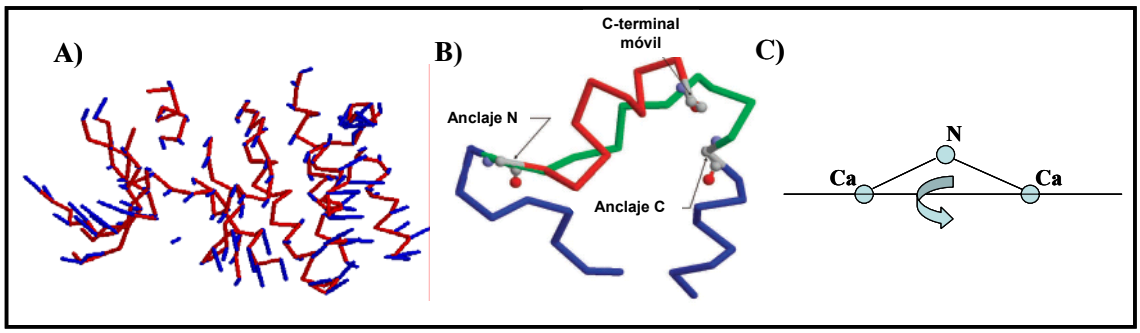


Figura 16. Resumen de los tipos de movimientos realizados en la simulación de estructuras. (A) Para el *centro estructural* se realizan movimientos aleatorios siguiendo las direcciones del espacio EPA; y para los *lazos*, (B) movimientos locales para el cierre de los mismos con CCD y (C) movimientos tipo *spike*. Además, también se realizan rotaciones locales de los extremos N- y C- terminales de la cadena proteínica.

Las regiones de *centro estructural* y de *lazos* se simulan por separado. En primer lugar, se optimiza el *centro estructural* realizando movimientos de los Cα a lo largo de las direcciones del *espacio EPA* (como se explicó en el apartado 3.8.2), y una vez alcanzada la conformación deseada, se optimizan los *lazos*. Para ello, además de los Cα también se consideraron los átomos de N de la cadena principal. La construcción de los *lazos* se inicia partiendo de estructuras al azar obtenidas mediante asignación de valores aleatorios a sus ángulos Cα-N-Cα desde sus extremos fijos N-terminales hasta conectar sus extremos libres C-terminales con sus puntos de anclaje C pertenecientes al *centro estructural* (ver Figura 14). A continuación, se van

escogiendo al azar residuos y se aplican tres tipos de movimientos dependiendo del tipode *lazo* en el que se encuentre el residuo escogido. Si el residuo pertenece al extremo N- o C-terminal de la cadena de proteína (considerados aquí como *lazos* a efectos de la simulación, aunque no lo sean en sentido estricto), se aplica el primer tipo de movimiento, consistente en rotar un ángulo al azar el segmento de *lazo* comprendido entre el residuo escogido *i*, y el extremo terminal, alrededor del enlace N_i-$C\alpha_i$. Si el residuo *i* pertenece a un lazo de menos de 4 residuos, se rota el átomo N_i alrededor del eje $C\alpha_i$-1-$C\alpha_i$, lo que constituye el segundo tipo de movimiento. Si el residuo *i* pertenece a un *lazo* de más de 4 residuos, se escoge al azar un segmento de 4 residuos en el *lazo* (*i*, *i*+1, *i*+2, *i*+3), y se varían ligeramente las posiciones de los átomos en el residuo (*i*+3) cambiando el ángulo N_{i+3}-$C\alpha_{i+3}$-N_{i+4} y rotando alrededor de la dirección $C\alpha_{i+3}$-N_{i+4}. Esta rotación se limita al intervalo [1.3 - 2.7] radianes, dada la restricción debida al carácter doble del enlace peptídico. Finalmente se aplica CCD para ajustar el segmento (*i*, *i*+1, *i*+2, *i*+3).

Al final de cada paso de la simulación, se evalúa la estructura del *lazo* resultante para determinar si es aceptable o no y el proceso se itera hasta convergencia.

3.8.5. Evaluación de la eficiencia del muestreo conformacional

En la Figura 17 se muestra un esquema del protocolo global de generación de modelos. Partiendo de la estructura problema, se realiza una búsqueda de homólogos con MAMMOTH de pares en la base de datos de estructuras. A continuación, se realiza un alineamiento estructural múltiple de todos los homólogos encontrados para detectar el *centro estructural conservado* y se estudia su plasticidad mediante análisis de componentes principales, lo que permite obtener el *espacio evolutivo* de deformación (*espacio EM-PCA*). Por otro lado, se calcula la estructura promedio de todos los homólogos encontrados para la diana, se le aplica un análisis de modos normales y seleccionando los de más baja frecuencia se obtiene el *espacio vibracional* (*espacio ANM*), que contiene las principales direcciones de deformación de la estructura debidas a su propia topología. A continuación se mezclan y ortogonalizan ambos espacios para obtener el *espacio de muestreo EPA*, cuyo origen de coordenadas se sitúa en la estructura promedio del conjunto de homólogos. Por último se realiza la simulación de intercambio de réplicas en este subespacio de muestreo y se selecciona la mejor estructura posible de acuerdo a una función de energía establecida.

Se permiten movimientos tanto del *centro estructural* de la proteína como de los *lazos*, de manera que se puede modelar la cadena completa, como se ha explicado en el apartado anterior. En una primera prueba se evaluó la eficiencia del muestreo usando una función de energía muy simple, consistente únicamente en el valor del RMSD entre la estructura muestreada y la problema (la estructura real de la proteína diana) y se obtuvieron resultados muy satisfactorios para las mejores estructuras muestreadas (ver apartado de Resultados 4.3, pág. 98).

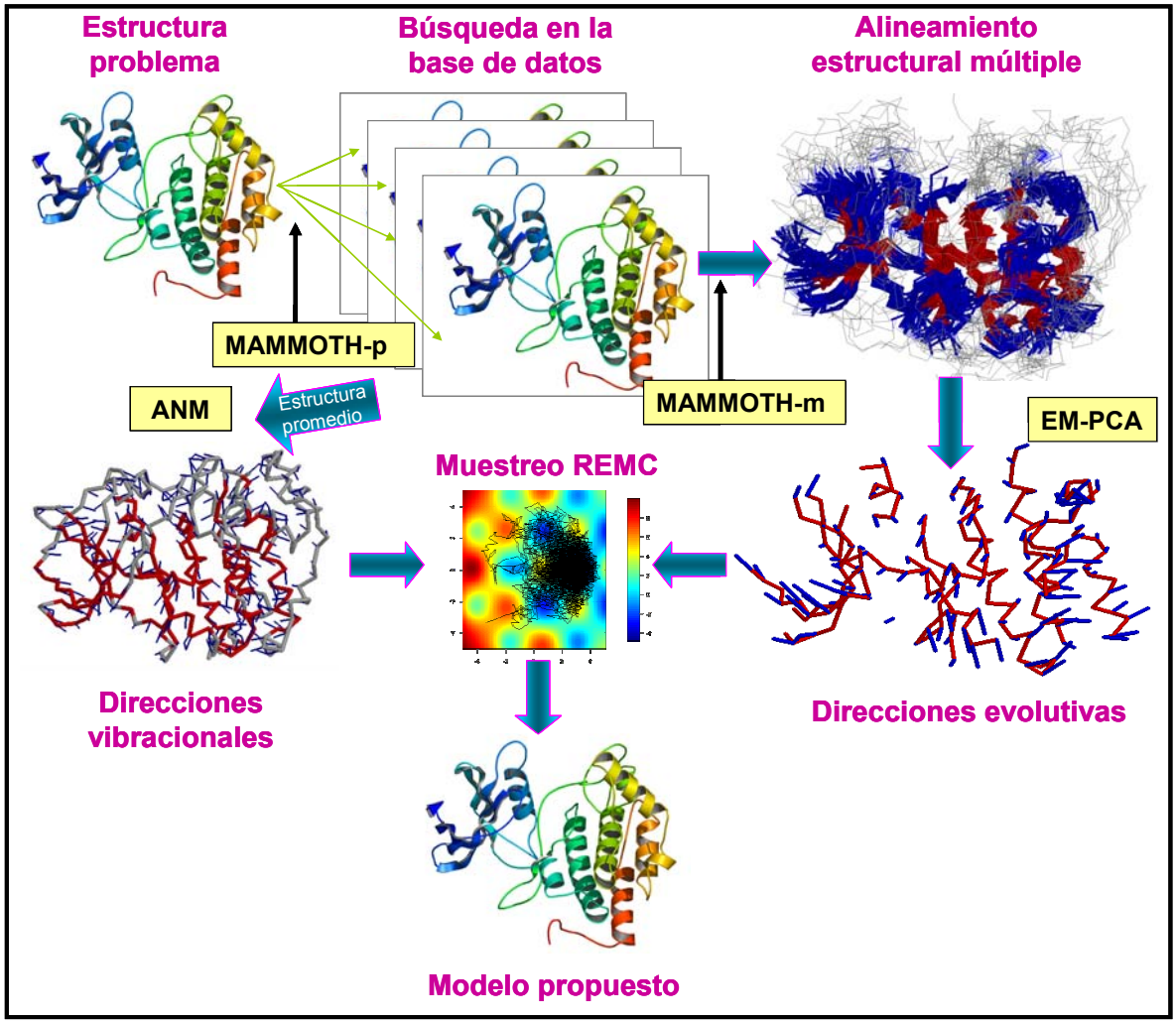


Figura 17. Esquema del protocolo general de generación de modelos.

No obstante, puesto que en una situación real no se dispone de la estructura diana, el perfil de energía creado por la función basada en el RMSD no es válido. Es por ello que además se exploró el comportamiento del método EPA-REMC bajo condiciones más realistas, creando perfiles energéticos de varios grados de dificultad y midiendo la capacidad del algoritmo de muestreo para alcanzar el mínimo global.

En esta evaluación, la función de energía incluye tres términos:

$$E(\mathbf{c}) = \sum_{i=1}^{nprot} h_i \exp\left[-\frac{(\mathbf{c}-\boldsymbol{\alpha}_i)^2}{2w_i^{\,2}}\right] + \sum_{i=1}^{ruido} p_i \exp\left[-\frac{(\mathbf{c}-\boldsymbol{\beta}_i)^2}{2\sigma_i^{\,2}}\right] + k(\mathbf{c}-\mathbf{c}_0)^2 \qquad (51)$$

donde **c** es la posición muestreada de la cadena en el espacio EPA de 50 dimensiones.

El primer término de la energía está compuesto por funciones gaussianas centradas en las estructuras de las proteínas de la familia. Su función consiste en simular atractores en los puntos correspondientes a las estructuras reales de la familia, *nprot* es el número de proteínas, h_i y w_i son respectivamente, la altura y la anchura de los picos gaussianos centrados en la proteína *i*, y

$\boldsymbol{\alpha}_i$ es la posición de la proteína *i*. Todos los valores h_i tienen los mismos valores negativos excepto h_{idiana} el cual, como su propio nombre indica, pertenece a la diana y se escoge más pequeño que los demás.

En la segunda parte, se introduce el ruido en la función de energía. Esta parte también está compuesta por funciones gaussianas, donde *ruido* es el número de picos gaussianos; p_i y σ_i son la altura y la anchura de la función gaussiana i, y $\boldsymbol{\beta}_i$ es el centro del pico gaussiano *i*. El parámetro de altura, p_i se escoge aleatoriamente con valor negativo o positivo, en el rango de $[0.8p_m, p_m]$ o $[-p_m, -0.8p_m]$ donde p_m es un parámetro positivo. Este término se introduce para reflejar los errores existentes en la función de energía. Los picos de ruido se distribuyen regularmente en el *espacio EPA*. Los parámetros de altura p_i para picos gaussianos vecinos se escogen con signos opuestos, de manera que todos los mínimos de energía están rodeados de barreras para incrementar la dificultad del muestreo. El tercer término de la función de energía es un término de sesgo general para crear una forma global del perfil energético de tipo-embudo, donde $\mathbf{c}_0$ corresponde a las coordenadas de la diana, y $k = 0.001$ es el parámetro que controla la pendiente del embudo. El valor k usado es pequeño, con el fin de hacer el muestreo más difícil.

La rugosidad de la superficie creada se puede evaluar mediante la diferencia de energía entre el mínimo global y el pico de ruido más bajo y el *Z-score*, que se define como:

$$Z_{score} = \frac{E_{\min} - < E >}{\sigma} \tag{52}$$

donde $E_{\min}$ es el mínimo global de energía, y $< E >$ y σ son la media y la desviación estándar de la función de energía en la superficie completa. Como puede apreciarse, la rugosidad de la superficie se puede modificar fácilmente cambiando P_m. El salto de energía se puede representar como $E_{salto} = P_m - h_{idiana}$ suponiendo que la diana corresponde al mínimo global de energía. Aparte de explorar la capacidad del método REMC para alcanzar el mínimo global en las diferentes superficies de energía creadas, se comparó también su comportamiento con respecto a otros métodos comúnmente usados en búsqueda conformacional: SA (S*imulated Annealing*) y RS (*Random Search*), (ver apartado de Resultados, pág. 100).

Se muestra la evaluación de la eficiencia del muestreo en 4 superficies diferentes con Z_{score} = -3.437, -3.221, -3.395 y -3.267, que corresponden a valores representativos de funciones de energía actualmente utilizadas en predicción de estructuras. En todos los casos, lo único que se cambió es el parámetro de altura P_m = 4.5 y 3.5 y el número de picos de ruido *ruido* = 2120, 2903, dejando fijos los otros parámetros. La altura para el pico de la diana se estableció en h_{idiana} = -5.00. El punto de partida de la simulación en los dos primeros autovectores se situó en el punto [-3.34 Å, -3.34 Å], a 7.19 Å de la diana (cuyas coordenadas en el mismo sistema 2-D

fueron [3.05 Å, 0.73 Å]). Se simularon 8 réplicas (o temperaturas), ya que el número de grados de libertad es 50 (el número de dimensiones del espacio EPA), siendo las temperaturas de simulación las siguientes: 2.000, 0.636, 0.313, 0.186, 0.123, 0.087, 0.065 y 0.050. Para cada réplica, se intentó el intercambio entre ellas cada 10 pasos de MC, dando un total de 10^5 intentos de intercambio (ver apartado de Resultados, pág. 98).

RESULTADOS

4. Resultados

4.1. Nuevo método de alineamiento estructural múltiple, MAMMOTH-mult

Se desarrolló un nuevo programa de alineamiento estructural múltiple, MAMMOTH-mult (Lupyan et al., 2005) . Los alineamientos obtenidos con este nuevo método suponen una mejora sobre los métodos manuales o automáticos disponibles anteriormente en muchas bases de datos ampliamente usadas a todos los niveles estructurales. Un análisis detallado de los alineamientos estructurales producidos para unos cuantos casos representativos indica que MAMMOTH-mult produce árboles con significado biológico a nivel de secuencia y estructura de motivos funcionales en los alineamientos. Un avance importante de MAMMOTH-mult con respecto a otros métodos es la reducción en coste computacional. Normalmente, los alineamientos con MAMMOTH-mult no llevan más de 5 segundos de CPU en promedio en un procesador simple R12000. Esto hace que el método sea particularmente útil para aplicaciones a gran escala.

En este apartado, se compara la calidad de los alineamientos obtenidos con MAMMOTH-mult con la de otros métodos de alineamiento estructural de proteínas comúnmente usados, se detalla su comportamiento para dos casos representativos (inmunoglobulinas y globinas) y se presenta un servidor web desarrollado para dar la posibilidad de usar el método *on-line*.

4.1.1. Calidad del alineamiento estructural múltiple

El comportamiento del algoritmo de alineamiento estructural múltiple se resume en la Tabla 2.

Método de Alineamiento	*Nivel Estructural*	*Zscore (%centro)*	*Zscore (<$RMSD_{centro}$>)*	*Zscore (norMD)*
A) *Semi-manual*				
HOMSTRAD	Familia	1.03	0.31	0.10
CAMPASS	Superfamilia	9.17	8.78	4.81
B) *Automático*				
DaliLite	Familia	4.97	3.39	5.41
	Superfamilia	7.04	4.72	4.20
	Plegamiento	2.73	2.37	2.27
MultiProt	Superfamilia	11.05	5.22	7.60

Tabla 2. Comparación de la calidad de los alineamientos estructurales para diferentes conjuntos de datos. Se calculan los tres parámetros que describen la calidad del alineamiento (*%centro*, <$RMSD_{centro}$> y *norMD*), para cuatro conjuntos de datos (FSSP, MultiProt, HOMSTRAD y CAMPASS, ver Métodos, pág. 27) usando tanto MAMMOTH-mult como un método de referencia (DaliLite y MultiProt para el caso de alineamientos automáticos, y alineamientos obtenidos a partir de los correspondientes servidores web para HOMSTRAD y CAMPASS). Se aplicó un test de Wilcoxon de suma de rangos a los dos pares de grupos. Se muestra el *Z-score* de la estadística resultante. Valores positivos indican una mejora de MAMOTH-mult sobre el método alternativo, y lo contrario para los valores negativos. Como se puede apreciar, en todos los casos se obtuvieron valores positivos. *Z-scores* > 3.0 indican diferencias estadísticamente significativas.

A continuación se discuten los resultados con respecto a los datos de referencia para los métodos semimanuales HOMSTRAD y CAMPASS. Para las 105 familias de HOMSTRAD, los valores promedio de los tres parámetros de puntuación usados para evaluar el éxito del método (ver apartado de Métodos, pág. 36), fueron los siguientes:

%centro = 71%, $<RMSD_{centro}>$=0.80 Å y *norMD* = 0.63 para MAMMOTH-mult y

%centro = 67%, $<RMSD_{centro}>$=0.78 Å y *norMD* = 0.63 para HOMSTRAD.

Como se aprecia, los resultados de MAMMOTH-mult fueron ligeramente mejores que los de HOMSTRAD pero las diferencias en los rangos no son estadísticamente significativas. Sin embargo, a medida que las estructuras dentro de la familia empiezan a divergir, MAMMOTH-mult tendió a dar mejores alineamientos estructurales que HOMSTRAD (Figura 18).

La Figura 19A y Figura 19B suponen un ejemplo extremo. En ellas se muestran las superposiciones obtenidas por HOMSTRAD y MAMMOTH-mult para 8 miembros de la familia de lectinas tipo-C. Para ellas, el alineamiento HOMSTRAD dio el 19.82% de los residuos en el *centro estructural* y un $<RMSD_{centro}>$ = 1.01 Å, mientras que con MAMMOTH-mult el 59.50% de los residuos estaban en el *centro* y tenían un $<RMSD_{centro}>$ = 0.86 Å. La tendencia de MAMMOTH-mult de generar mejores alineamientos, comparado con los métodos manuales, con estructuras más divergentes, se confirma con el conjunto de CAMPASS. En este caso (Figura 18), MAMMOTH-mult da mejores alineamientos, con diferencias fuertemente significativas (Tabla 2), particularmente en cuanto al tamaño del *centro* y a la fluctuación de RMSD de los residuos del mismo.

La Figura 19C y Figura 19D muestran un ejemplo de las diferencias en el alineamiento para la superfamilia de 8 miembros de ribonucleoproteínas Sm. CAMPASS dio un alineamiento final con el 6.12% de residuos en el *centro* y un $<RMSD_{centro}>$ = 1.84Å, mientras que con MAMMOTH-mult, el alineamiento tenía el 40.62% de residuos en el *centro* y un $<RMSD_{centro}>$ = 1.44 Å

Cuando se compararon los alineamientos de MAMMOTH-mult con otros métodos automáticos, en la comparación con MultiProt usando el conjunto de alineamientos de CAMPASS (ver apartado de Métodos, pág. 27), se observó una mejora significativa para los alineamientos con MAMMOTH-mult (ver Tabla 2). Las mejoras son particularmente significativas en el tamaño del *centro* detectado y en este caso, en la calidad del alineamiento de secuencia correspondiente. También se hizo una comparación con los alineamientos de FSSP, obtenidos con DaliLite. Para ello, primero se aseguró la compatibilidad entre ambos programas (como se ha explicado en el apartado de Métodos, 27). Para los 1385 alineamientos seleccionados, los valores promedio de los tres parámetros de puntuación fueron los siguientes:

%*centro* = 37%, $<RMSD_{centro}>$=1.01 Å y *norMD* = -0.89 para MAMMOTH-mult y

%*centro* = 33%, $<RMSD_{centro}>$=1.18 Å y *norMD* = -1.19 para FSSP.

A) HOMSTRAD **B) CAMPASS**

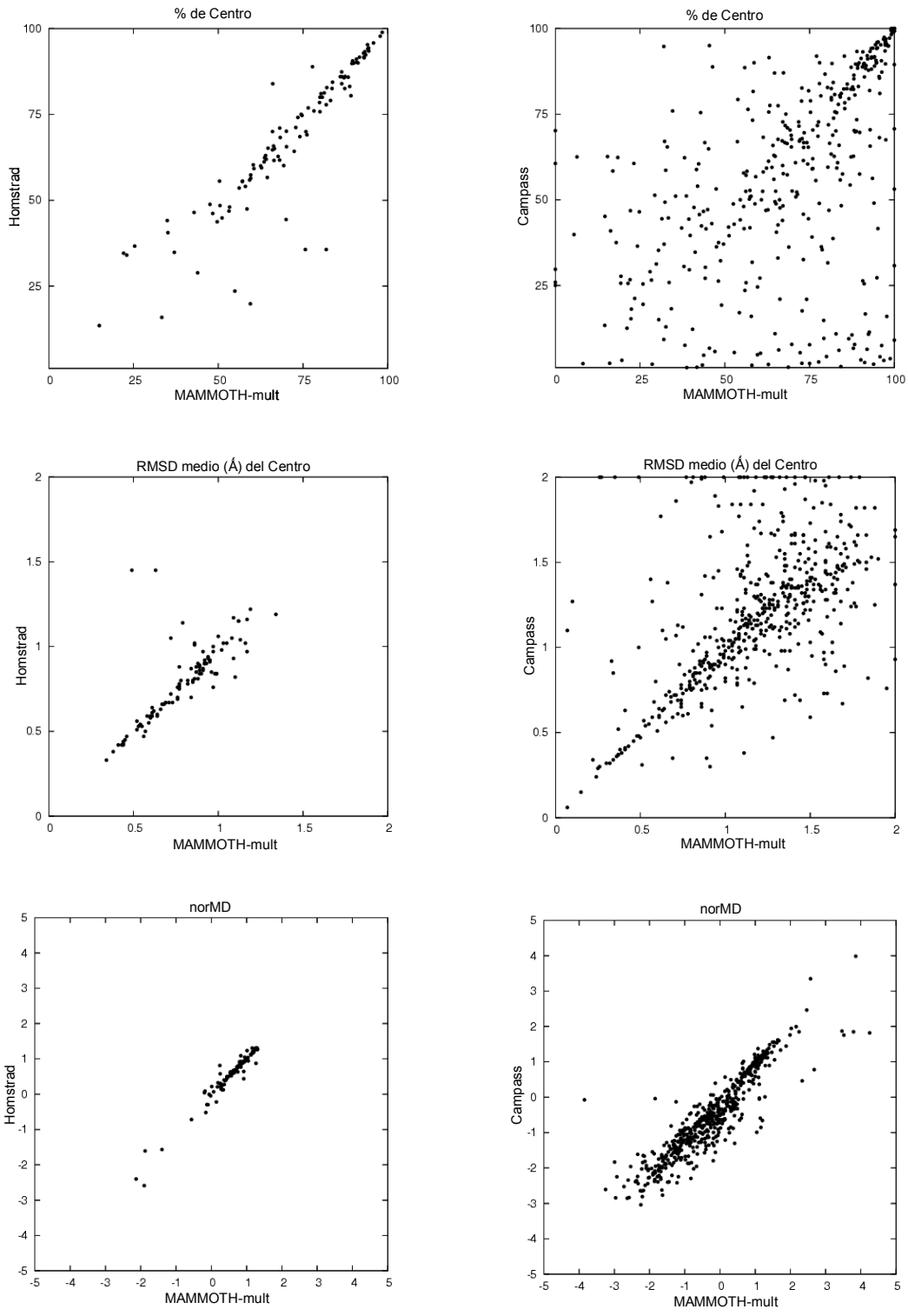


Figura 18. Alineamientos estructurales generados con MAMMOTH-mult en comparación con otros métodos. El eje "x" se refiere a los resultados de MAMMOTH-mult, mientras que el "y" se refiere a los del método de referencia. Cada punto corresponde a una familia de proteínas, con un número promedio de 7 miembros por familia. Se muestran gráficos para el porcentaje de centro (%*centro*), fluctuación media del RMSD de los residuos en el *centro* (<$RMSD_{centro}$>), y valor *norMD*. A) HOMSTRAD; B) CAMPASS.

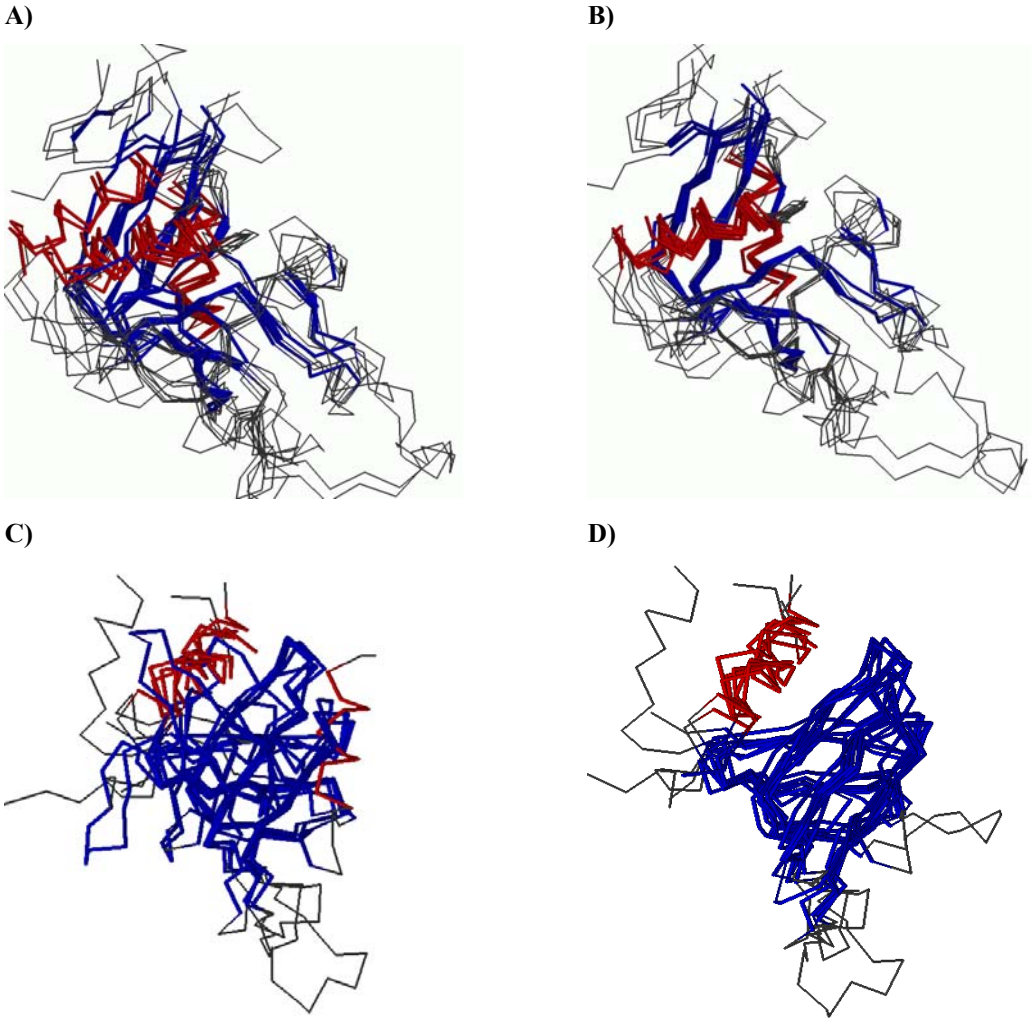


Figura 19. Alineamientos estructurales para los 8 miembros de la familia de lectinas tipo-C (A, HOMSTRAD; B, MAMMOTH-mult), y los 8 miembros de la superfamilia Sm (C, CAMPASS; D, MAMMOTH-mult). Las regiones coloreadas (azul para las láminas y rojo para las hélices), resaltan las regiones del *centro* estructuralmente conservadas, según se definen en MAMMOTH-mult.

Cuando los alineamientos se dividen en clases estructurales (*familia*, *superfamilia* y *plegamiento*, ver Tabla 2), se observaron grandes mejoras a nivel de *familia* y particularmente a nivel de *superfamilia*, pero más modestas a nivel de *plegamiento*. Hay que resaltar que en muchos de los alineamientos de FSSP, así como en algunos de MAMMOTH-mult, no fue posible detectar *centro estructural* convencional (0% de *centro*); estos casos no se consideraron en el cálculo de las medias.

Finalmente, también se llevaron a cabo alineamientos estructurales múltiples con dos conjuntos de estructuras usadas por Ochagavia y Wodak en su validación de MALECON (Ochagavia and Wodak, 2004). El primero consistió en un conjunto de **globinas**, formado por las siguientes estructuras: 1ash, 1eca, 1gdj, 1hlm, 1mba, 1babA, 1ew6A, 1h97A, 1ithA, 1sctA, 1dlwA, 1flp, 1hbg, 1lhs y 1vhbA. Aunque la comparación directa de resultados se debe hacer con precaución, dada la ligeramente distinta definición de *centro estructural*, el *centro* detectado

por MAMMOTH-mult abarcó 131 residuos con un $<RMSD_{centro}>$= 1.56 Å, mientras que el *centro* encontrado por MALECON contuvo 59 residuos con un $<RMSD_{centro}>$= 1.73 Å.

Para el segundo caso, el conjunto **OB** (1afp, 1b9nA3, 1ckmA2, 1esfA1, 1fr3A, 1jic, 1tiiD, 2tmp, 1b7yB2, 1bovA, 1eif02, 1fjgQ, 1htp, 1sro y 2sns), ambos métodos mostraron un comportamiento similar (fallaron al encontrar un alineamiento con el conjunto completo). Cuando este conjunto se redujo a 10 estructuras (1sro, 1b7yB2, 1tiiD, 1bovA, 2sns, 1esfA1, 1eif02 1fjgQ 1b9nA3 y 1fr3A), MAMMOTH-mult y MALECON también produjeron resultados similares (datos no mostrados).

4.1.2. Análisis de casos representativos

Se analizaron los alineamientos derivados automáticamente por MAMMOTH-mult para dos casos bien conocidos. En nuestro grupo se hicieron comparaciones extensivas para un gran número de casos, pero por motivos de espacio, sólo se muestran aquí resultados para dos ejemplos representativos (descritos como el conjunto de *Superplegamiento* en el apartado de Métodos, pág. 27): inmunoglobulinas y globinas.

Inmunoglobulinas

Bork et al. (Bork et al., 1994), clasificaron semi-manualmente 26 dominios diferentes de inmunoglobulinas en cuatro subtipos diferentes (tipo-v, tipo-h, tipo-s y tipo-c). Su estudio agrupó las estructuras en función del número, conexión y variaciones en la posición relativa de las hojas beta de los bordes con respecto a un *centro estructural* común de cuatro hojas beta. El *centro* común encontrado por MAMMOTH-mult se puede observar en la Figura 20, usando la notación empleada por Bork et al.

Se observó que MAMMOTH-mult también encontró un *centro* común de cuatro láminas centrales bien conservadas (B, C, E y F), y variaciones en las posiciones de las láminas de los bordes A, C' y G (Figura 20), y de acuerdo con su estudio, una considerable variación en las láminas A, C'' y D, que no forman parte del *centro evolutivo* de MAMMOTH-mult.

Sin embargo, el cuarto grupo (tipo-h), sólo consistió en una estructura (1gof), con los otros dos miembros (1cgt y 1clc) distribuidos entre los otros tipos c y s. Esto no resultó sorprendente, ya que el grupo tipo-h es considerado por Bork et al. un híbrido entre ellos. En cuanto al alineamiento de secuencias correspondiente, se observó una conservación clara de los residuos aromáticos en las láminas del *centro*, particularmente para las C y F (Figura 20), característica también observada por Bork et al. De manera similar, se detecta un incremento en la longitud del segmento CD al ir desde el tipo s al c y al v, como también fue apuntado en el trabajo de Bork.

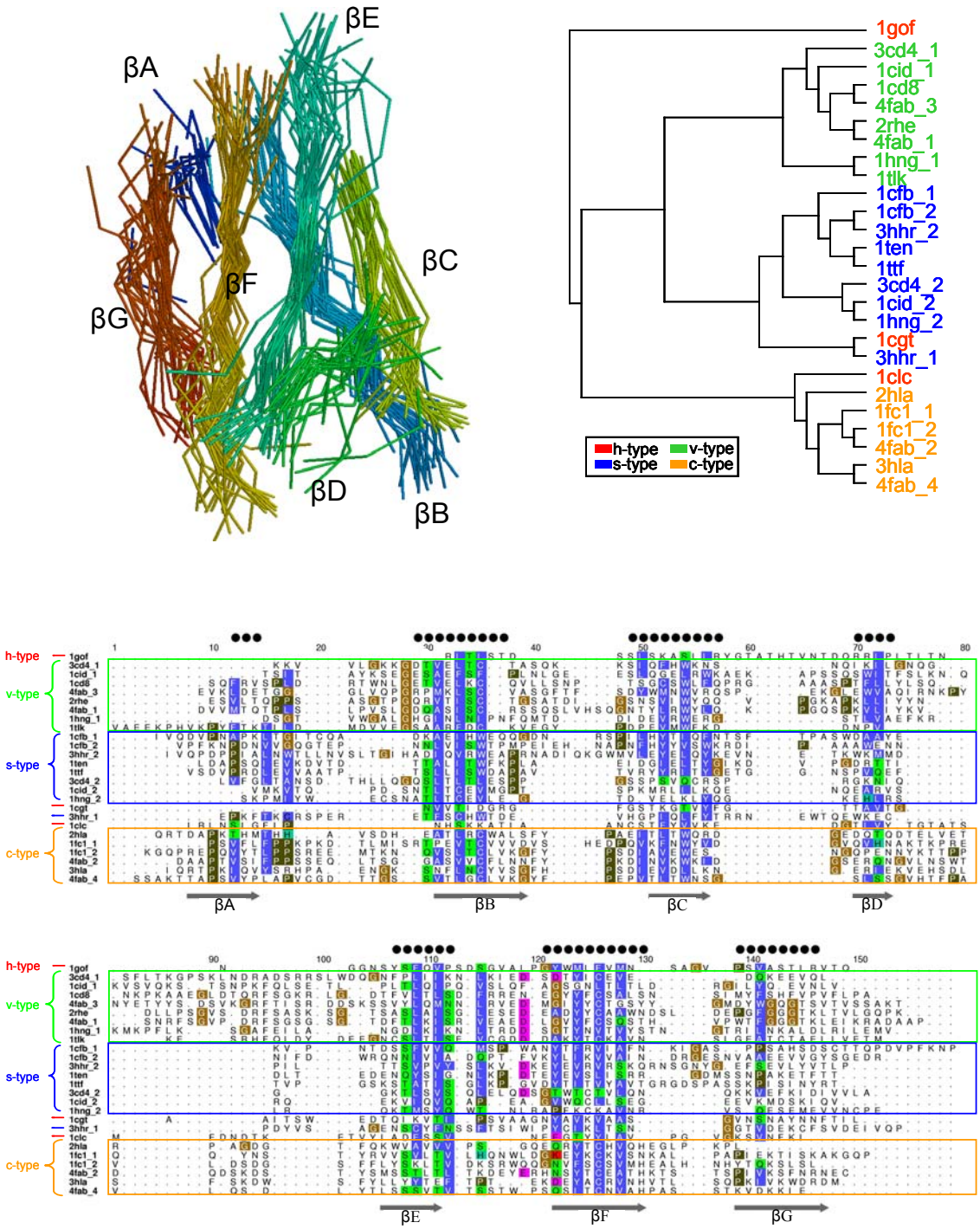


Figura 20. Alineamiento estructural de MAMMOTH-mult para las Inmunoglobulinas. Se muestran el alineamiento múltiple de secuencias correspondiente, el alineamiento structural para el *centro evolutivo* (“*centro permisivo*”, ver Métodos, pág. 36), calculado con MAMMOTH-mult y el dendograma correspondiente a ese alineamiento estructural. Los nombres de las estructuras están coloreados en el dendograma de acuerdo a la clasificación manual de referencia (Bork et al., 1994).

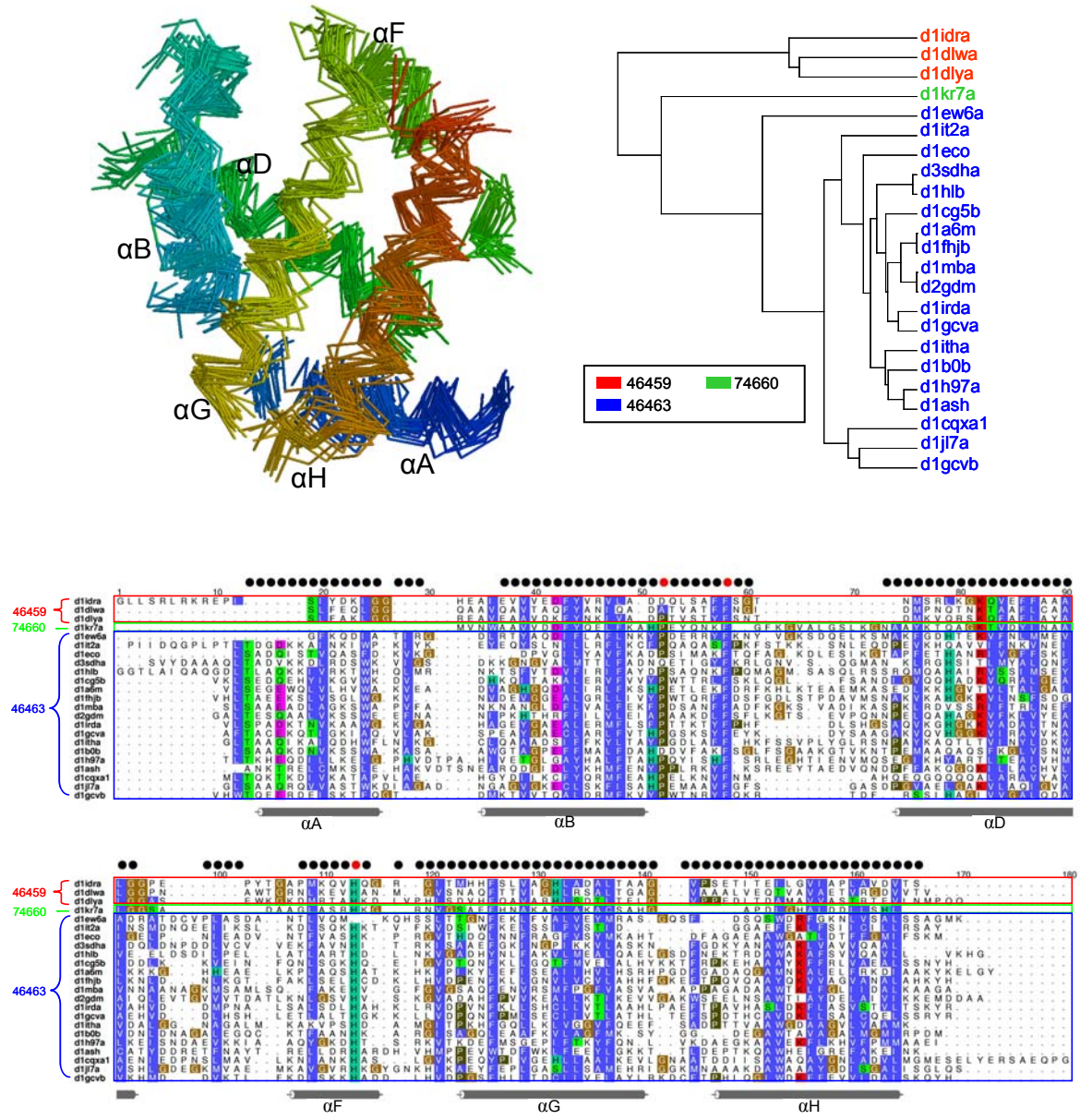


Figura 21. Alineamiento structural de MAMMOTH-mult para las Globinas. Se muestran el alineamiento múltiple de secuencias correspondiente, el alineamiento structural para el *centro evolutivo* ("*centro permisivo"*, ver Métodos), calculado con MAMMOTH-mult y el dendograma correspondiente a ese alineamiento estructural. Los nombres de las estructuras están coloreados en el dendograma de acuerdo a la familia a la que pertenecen según SCOP.

Globinas

Las globinas se organizan en el llamado "*plegamiento tres-en-tres α-helicoidal*" en SCOP. Un conjunto de 23 globinas (con entrada de SCOP 46458 para el nivel de superfamilia), se seleccionó de ASTRAL de manera que todas las identidades en secuencia por pares entre ellas eran menores del 40%. Este conjunto estaba formado por tres familias de SCOP: 46459, 74660 y 46463. Se observó que el árbol producido por MAMMOTH-mult reproducía exactamente la clasificación de SCOP, separando globinas canónicas hemo-enlazantes (46463) de las hemoglobinas truncadas de protozoos y bacterias (46459) y las hemoglobinas neuronales (74660) (Figura 21). En el alineamiento de secuencias correspondiente, se observó una clara conservación de la histidina proximal clave en el sitio activo (posición 113, HisF8 siguiendo la notación de Perutz (Perutz, 1960)). Esta histidina establece un enlace Fe-N con el átomo de Fe en el grupo hemo y su conservación es una de las características distintivas del perfil de globinas (Kapp et al., 1995). La única proteína que falló al alinear la histidina en esta posición fue la d1ew6a_ (LaCount et al., 2000), de *Amphitrite ornate*. Resultó interesante que fuera la única deshaloperoxidasa del conjunto. Para esta proteína, la histidina proximal estaba desplazada tres posiciones. Probablemente, este desplazamiento fuerce una rotación de 60º en el imidazol, lo que concordaría con estudios previos que han sugerido que esta rotación podría ayudar a establecer un fuerte enlace Fe-N que contribuiría al flujo de electrones requerido por la actividad peroxidasa (LaCount et al., 2000). También se observó una conservación absoluta de la fenilalanina PheCD1 (posición 58), considerada un residuo clave en la interacción de la proteína con el grupo hemo, y totalmente conservada también en los alineamientos de globinas (Kapp et al., 1995). Finalmente, la ProC2 (posición 52), en la vecindad inmediata del bolsillo enlazante a hemo, también estaba casi completamente conservada, en concordancia con análisis previos (Ptitsyn and Ting, 1999).

4.1.3. Tiempos de ejecución

Los tiempos computacionales son cruciales para aplicaciones a gran escala. En la Figura 22 se muestra la dependencia del tiempo de computación cuando se ejecuta MAMMOTH-mult con respecto al número de residuos a alinear. Un alineamiento típico de 15 estructuras con 150 residuos cada una, tarda unos 5 s usando un Pentium IV PC a 2 GHz y unos 25 s en un R12000. Alinear el conjunto completo de 105 familias estructurales de HOMSTRAD requirió 27 minutos de CPU en un solo procesador R12000. Como dato de comparación, el método de Nussinov et al. (Leibowitz et al., 2001) tarda unas 10 h de tiempo de CPU sólo para alinear 10 TIM barrels. Con MAMMOTH-mult, el mismo conjunto de estructuras se pudo alinear en 28 s en un procesador R12000. Las razones de estos bajos tiempos de computación son: 1) el algoritmo de alineamiento por pares subyacente en el que se basa esta versión múltiple, ya es de por sí muy

rápido y 2) en cada nodo en los pasos 3.1 y 3.2 del algoritmo múltiple (ver apartado de Métodos, pág. 29), el *URMS* se obtiene de tablas de registro que se llenan durante las comparaciones por pares hechas en el paso 1.

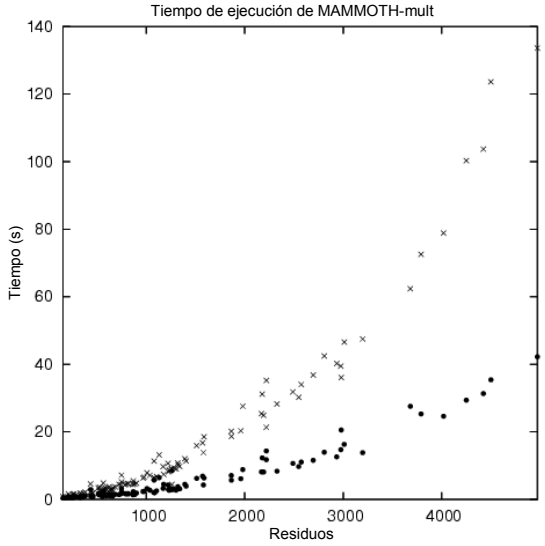


Figura 22. Tiempos de ejecución de MAMMOTH-mult. Se representa el número total de residuos alineados en la familia frente al tiempo de ejecución (en segundos). Se muestran los resultados para los cálculos en un procesador R12000 (cruces) y en un PC Pentium IV a 2 GHz con sistema operativo Linux Red-Hat (puntos).

4.1.4. Servidor web para uso de MAMMOTH-mult

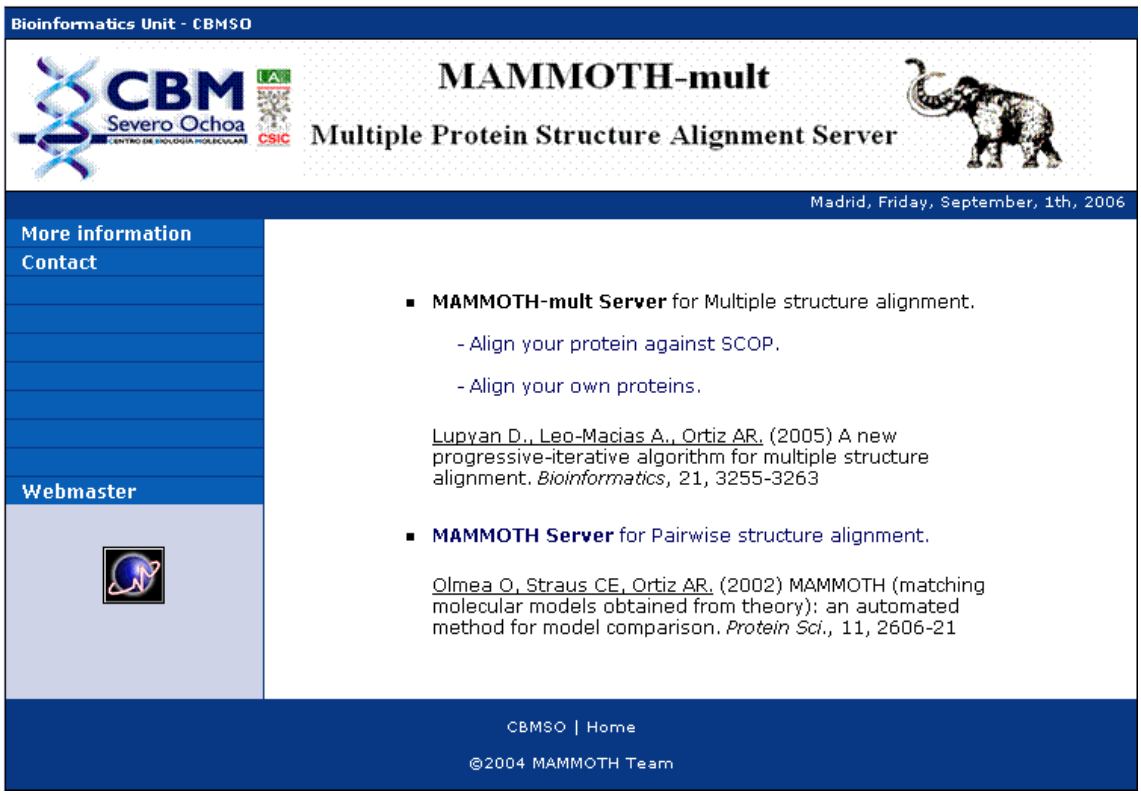
Para permitir el uso *on-line* del programa, se implementó un servidor web en la siguiente dirección: http://ub.cbm.uam.es/mammoth/mult (ver Figura 23). El servidor se puede usar de dos maneras diferentes: bien alineando una proteína problema contra una superfamilia dada de SCOP, o alineando entre ellas un conjunto de proteínas de entrada. En el primer caso, se presenta al usuario un formulario donde se puede subir la estructura de una proteína en formato PDB y un índice para la superfamilia de SCOP contra la que se quiere alinear. El servidor realiza una consulta automática a la base de datos y muestra todos los dominios pertenecientes a la superfamilia elegida. El usuario debe entonces seleccionar un subconjunto de los dominios mostrados para llevar a cabo al alineamiento. En el segundo caso, el fichero de entrada a subir consiste en un único archivo con todas las proteínas a alinear concatenadas una detrás de otra en formato PDB y separadas con el término "TER". En ambos casos el servidor lleva a cabo el alinemiento y envía los resultados por correo electrónico al usuario. Se devuelven seis archivos:

1. ***file*.log:** Contiene información acerca del alineamiento (número de proteínas, número de residuos, *z-score*, etc…).
2. ***file*.pdb:** Contiene las coordenadas de las proteínas superpuestas en formato PDB.

3. ***file*.tcl:** Contiene un *script* para Rasmol (Sayle and Milner-White, 1995) que permite una representación más detallada del archivo file.pdb, con diferentes colores y grosores para aquellos residuos que se quieren resaltar.
4. ***file*.aln:** Contiene el alineamiento de secuencias en formato ALN, derivado del alineamiento estructural.
5. ***file*.rot:** Contiene las matrices de traslación y rotación calculadas para cada proteína del alineamiento.
6. ***file*.dnd:** Contiene el dendograma de las estructuras del alineamiento en formato NEWICK.

También se implementó una interfaz web para el uso *on-line* de la versión de pares de este método de alineamiento estructural. En este caso, como entrada, se requieren dos archivos separados que contienen las coordenadas en formato PDB de las dos proteínas a alinear. El servidor realiza el alineamiento estructural y como salida, devuelve tres archivos: ***file*.log**, ***file*.pdb** and ***file*.tcl**.

A)



B)

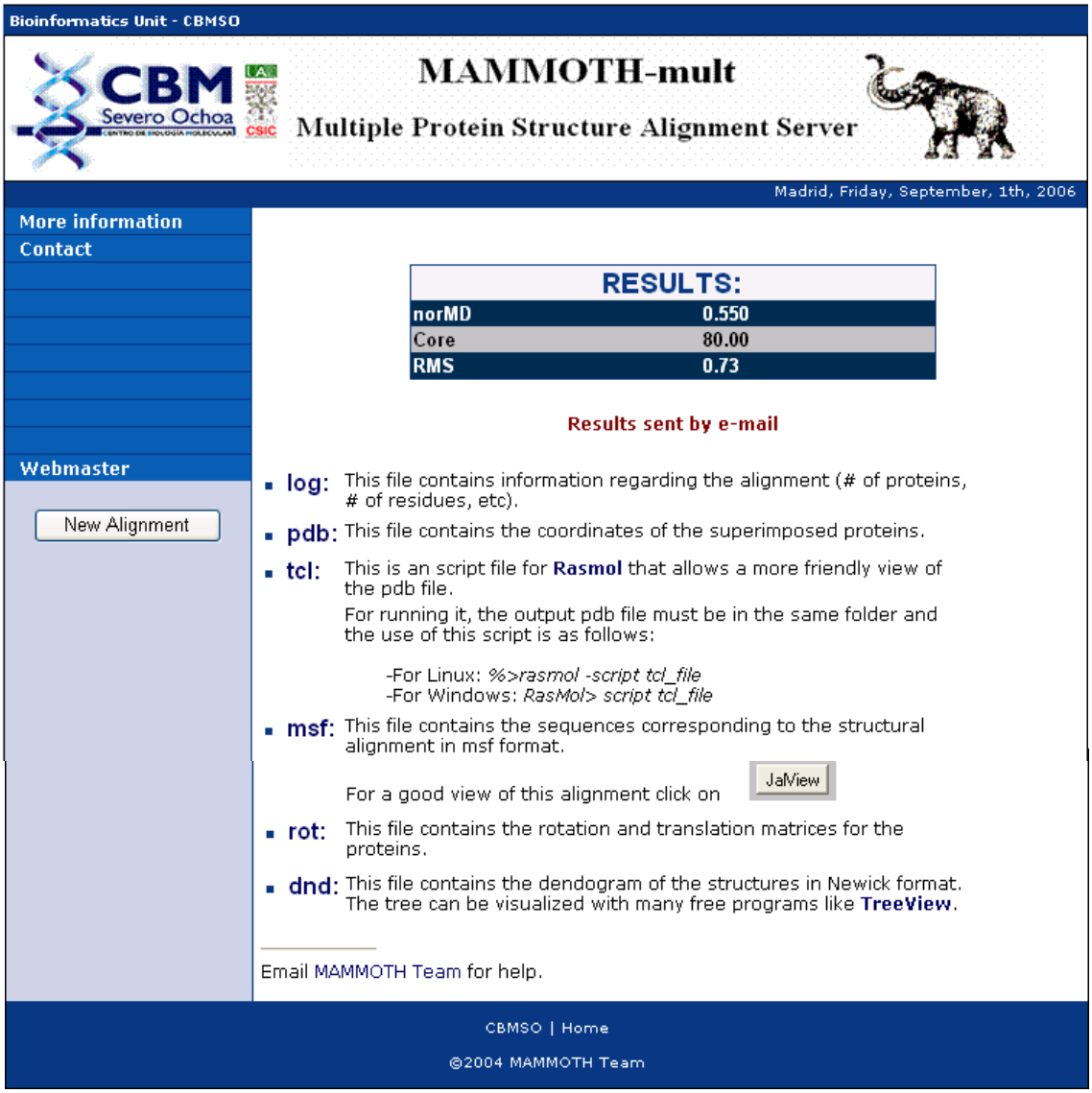


Figura 23. Imágenes de las páginas de entrada y salida del servidor. A) Acceso tanto a la versión de pares de MAMMOTH como a la versión múltiple. B) Ejemplo de salida para un alineamiento múltiple, con enlaces a los archivos generados y los valores resultantes de *norMD*, *% de centro estructural* detectado y *RMSD* en Å.

4.2. Estudio del espacio de muestreo, EPA

4.2.1. PCA

Se estudió la plasticidad de los *centros estructurales* de las familias de proteínas homólogas. Para ello, cada una de las 35 superfamilias del conjunto de datos 3.1.2 (ver Tabla-Mat.Sup. 1), se sometió a alineamiento estructural múltiple con MAMMOTH-mult para caracterizar su *centro estructural* y se aplicó un análisis de componentes principales para estudiar las deformaciones más importantes que habían tenido lugar en él a lo largo de la evolución (ver en Tabla 3 un resumen de los resultados).

El *centro estructural* detectado en estos alineamientos y usado posteriormente en el análisis de componentes principales, PCA, contuvo en promedio un 62.4 ± 12.5% del total de la estructura (porcentaje tomado con respecto a la proteína más corta de la superfamilia). Por otra parte, la desviación cuadrática media de este *centro estructural* fue RMSD=2.07 ± 0.60 Å. Tanto el número de estructuras usadas en el alineamiento como el tamaño de los *centros* detectados fueron lo suficientemente grandes como para asegurar que las deformaciones detectadas usando PCA se aproximaban a las deformaciones reales experimentadas por la familia de proteínas.

	# Estrct.	*# Centro res.*	*# PC´s 70% var.*	*<Rmsd> ± dev*	*% Centro*
GLOBINAS	23	75	5	1.89 ± 0.63	69
QUINASAS	22	166	6	2.03 ± 0.47	64
INMUNOGLOBULINAS	23	50	6	1.92 ± 0.54	58
TRANSFERASAS S GLUTACION	22	67	6	1.90 ± 0.51	59
INTERLEUQUINAS	11	51	4	1.63 ± 0.71	83
DOMINIO DE UNIÓN A RNA	21	51	5	2.70 ± 0.59	68
FIBRONECTINAS	46	38	9	2.34 ± 0.89	45
CITOCROMO C	16	36	3	1.64 ± 0.43	46
TIOREDOXINAS	35	39	3	2.08 ± 0.84	53
SH3	24	34	5	1.87 ± 0.55	60
CUPREDOXINAS	22	48	4	2.00 ± 0.56	49
TOXINA DE SERPIENTE	11	36	4	1.49 ± 0.41	60
ALDOLASAS	19	84	5	2.07 ± 0.45	40
FERRITINA	15	103	4	1.97 ± 0.56	72
DOMINIO MUERTE	12	59	4	2.61 ± 0.62	71
RECEPTOR NUCLEAR DE UNIÓN A LIGANDO	14	175	6	1.92 ± 0.35	79
PECTIN-LIASA	11	111	4	2.08 ± 0.49	56
RIBOFLAVINA SINTASA	20	71	6	1.90 ± 0.38	78
LIPOCALINAS	23	62	4	2.18 ± 0.76	50
DOMINIO PDZ	15	56	6	1.94 ± 0.70	68
GAMMA-CRISTALINA	14	51	3	2.34 ± 0.97	67
DOMINIO LDH C-TERMINAL	12	114	3	2.04 ± 0.70	72
NTF2	14	83	4	2.35 ± 0.51	74
ABRAZADERA A ADN	11	69	3	2.35 ± 0.93	63
DOMINIO ATPASA DE HSP90 CHAPER.	13	69	4	1.82 ± 0.50	49
ACYL-COA-N-ACYLTRANSFERASAS	13	64	6	2.17 ± 0.50	47
BARRIL DE UNION A RIBULOSA-FOSFATO	18	125	5	2.32 ± 0.43	63
EXOPEPTIDASAS ZN-DEPENDIENTES	11	119	3	2.67 ± 0.80	44
PROTEINA DE UNION PERIPLASMICA	13	103	4	2.42 ± 0.60	40
FOSFATASAS II (FOSFOTIROSINA)	12	92	4	1.89 ± 0.61	64
FERREDOXINA REDUCTASA	22	92	4	2.08 ± 0.40	77
DOMINIO SCR	12	34	5	2.24 ± 0.88	59
DEFENSINA	21	22	4	2.33 ± 0.56	73
DEDOS DE ZINC C2H2 Y C2HC	21	20	4	1.57 ± 0.51	77
TOXINA DE ESCORPIÓN	16	20	4	1.77 ± 0.74	87

Tabla 3. Resumen de los resultados para las superfamilias del conjunto de datos 3.1.2 (ver apartado de Métodos, pág. 28 para más detalles).

En la Figura 24 se pudo observar que las deformaciones estructurales de los *centros evolutivos* cubren un espacio de baja dimensionalidad, ya que el 70% de la varianza total de las deformaciones de los mismos se puede explicar con 4.5±1.2 componentes en promedio. Además, el comportamiento de todas las familias en el PCA resultó muy similar, independientemente de la clase estructural, del tamaño, o del número de estructuras. Aunque el muestreo estructural es clave para la definición del subespacio PCA, y no se puede estar completamente seguro de haber cubierto todo el espacio accesible a una familia de proteínas dada, la similitud de los resultados en todos los casos, sugiere que las conclusiones obtenidas son robustas.

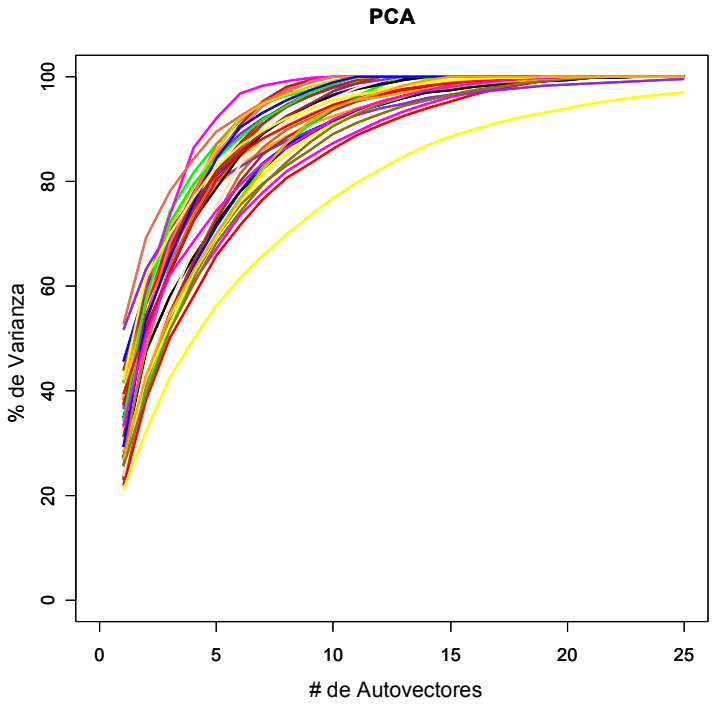


Figura 24. Porcentaje de la varianza explicada en función del número de autovectores del PCA.

Un hallazgo interesante también fue que el PCA resume las deformaciones evolutivas de una superfamilia en las direcciones que reflejan la mayoría de las adaptaciones funcionales. Se muestra un ejemplo en la Figura 25, que representa la distribución de estructuras de la superfamilia del receptor nuclear de unión a ligando (código 48508 de SCOP), en los dos primeros componentes principales. Se puede apreciar una clara separación funcional, a lo largo del primer componente que diferencia al grupo de los dominios de unión a esteroides (*grupo 1* en la figura), de los grupos de los dominios de unión a ácido retinoico y análogos (*grupo 2*).

Cuando se analizó el autovector correspondiente (Figura 26A), se vio que una de las regiones en la proteína que más contribuía al movimiento en ese autovector era el final de la hélice 5, lo que resultó muy interesante, ya que esta región incluía al residuo Arg-278, cuya

posición, en el sitio de unión al ligando, se sabe involucrada en la determinación de la selectividad (Steinmetz et al., 2001).

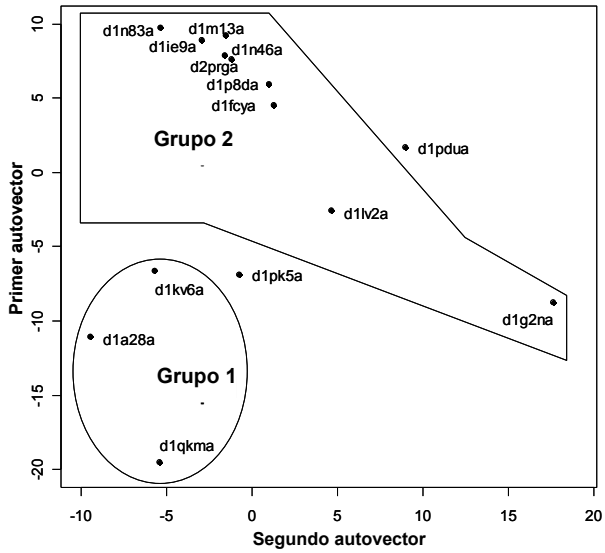


Figura 25. PCA de la superfamilia 48508 (dominio del receptor nuclear de unión a ligando). Se muestra la distribución de las estructuras en el plano formado por los dos primeros autovectores. El *grupo 1* corresponde a estructuras que reconocen ligandos esteroideos, mientras que el *grupo 2* corresponde a dominios que reconocen ácido retinoico y sus análogos. Las estructuras que no forman parte de ninguno de estos dos grupos, corresponden a receptores huérfanos.

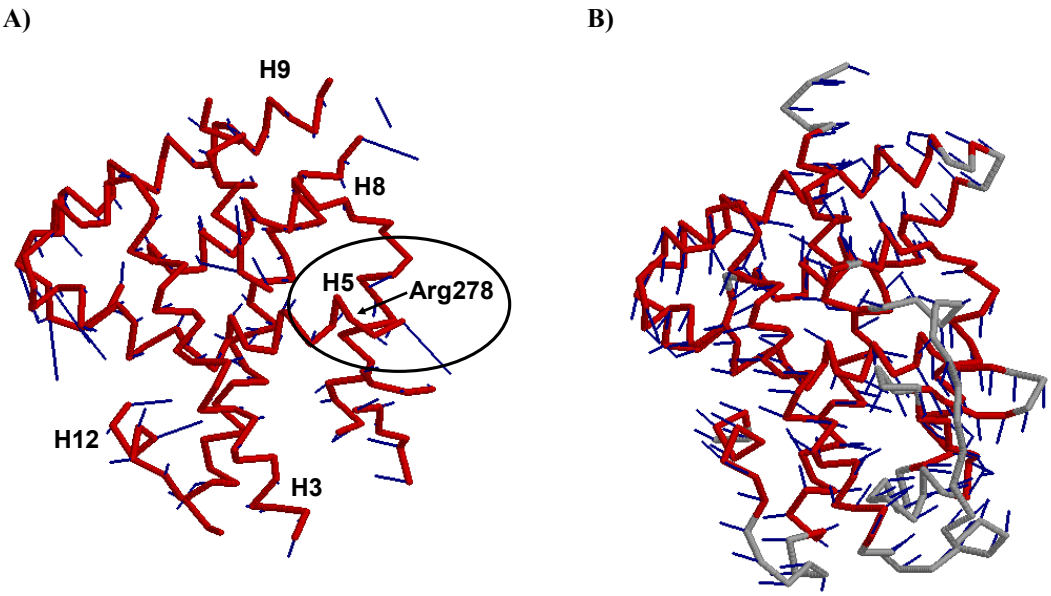


Figura 26. (A) *Centro estructural* promedio detectado por MAMMOTH-mult (traza de Cα en rojo), para la superfamilia 48508 (dominio del receptor nuclear de unión a ligando), y el primer autovector (segmentos en azul pegados a los residuos de la traza). También se señalan las diferentes hélices en la estructura. La contribución relativa de cada residuo al autovector viene dada por la longitud del segmento pegado al residuo. Se resalta especialmente el final de la hélice 5 (H5), que contiene el residuo Arg-278 implicado en la selectividad a ligando. (B) Se muestra el primer autovector de ANM. Los modos se calculan usando la estructura más cercana al promedio de la superfamilia (que se muestra en la figura).

4.2.2. Comparaciones PCA y ANM

Se estudió la relación entre los espacios PCA y ANM (ver Métodos, pág. 44). Para cada superfamilia, los cálculos de ANM se llevaron a cabo sobre aquella estructura más cercana a la estructura promedio determinada con MAMMOTH-mult. Consistentemente con estudios previos (Keskin et al., 2000), las pruebas indicaron que los modos normales no se ven significativamente afectados por la estructura específica de la superfamilia usada en el cálculo (no mostrado). Un ejemplo de un modo normal ANM se muestra en la Figura 26B, donde se representa el modo normal de más baja frecuencia calculado para un miembro representativo de la superfamilia de receptor nuclear de unión a ligando (48508), junto con la estructura empleada en el cálculo. La orientación de la estructura es la misma que la usada en la Figura 26A. Una inspección visual simple de las dos figuras, indica que los movimientos en ambos casos son considerablemente diferentes. Éste es generalmente el caso para la mayoría de las comparaciones por pares entre los autovectores del PCA y del ANM (no mostrado). Sin embargo, como se expone más adelante, se demostró que existe un subespacio dado de todo el espacio ANM completo que puede formar una base adecuada para representar los autovectores del PCA, incluso dada la pobre correlación encontrada entre los pares de autovectores.

Este hecho se pudo cuantificar midiendo la proyección de los autovectores PCA en el subespacio ANM, mediante el parámetro RMSIP (ver apartado de Métodos, pág. 45). Para ello, se restringieron las comparaciones usando los 50 primeros modos vibracionales de más baja frecuencia y el número de componentes principales del PCA mostrados en la Tabla 3 para cada superfamilia.

Primero se determinó el límite de distancia óptimo para seleccionar residuos vecinos en el modelo de red anisotrópico (ver apartado de Métodos, pág. 41). Para ello, se calculó el *Z-score* del RMSIP entre los espacios PCA y ANM para diferentes valores de corte, y se determinó que el valor óptimo para la distancia de corte entre dos residuos fue de 15 Å, dando un RMSIP medio de 0.85 (Figura 27). Este valor de distancia de corte era próximo al óptimo encontrado por Bahar et al., cuando compararon las fluctuaciones cuadráticas medias calculadas mediante ANM y las deducidas a partir de los B-factores experimentales (Atilgan et al., 2001). El valor de RMSIP obtenido resultó ser muy significativo, con un *Z-score* por encima de 15 (Figura 27B).

A continuación, para este valor de distancia de corte óptimo se estudió, cómo depende el solapamiento entre ambos espacios con el número de modos de baja frecuencia considerados, incluyendo hasta 50 modos. Los resultados se muestran en la Figura 28, en donde se representa el solapamiento en términos de *Z-score* promedio para las diferentes clases estructurales. Se observó que un solapamiento significativo se adquiría rápidamente dentro de los primeros 20 modos vibracionales y que a partir de ahí se tendía a un *plateau*. Las clases de proteínas pequeñas y α/β mostraron solapamientos significativamente más pequeños, mientras que las clases α y α+β presentaban los solapamientos más grandes. Para las proteínas pequeñas, la razón

del menor solapamiento encontrado podía ser el elevado número de puentes disulfuro que presentan (y que no son tenidos en cuenta en el ANM). En resumen, se puede decir que se encontró un solapamiento estadísticamente significativo entre las deformaciones observadas en el *centro estructural* de proteínas homólogas y los ~20 primeros modos de más baja frecuencia impuestos por la topología de la proteína. Por tanto, la estructura del *centro* de proteínas evolutivamente relacionadas, parece responder a los cambios en secuencia mediante deformaciones a lo largo de combinaciones de modos normales impuestos por la topología.

A) **B)**

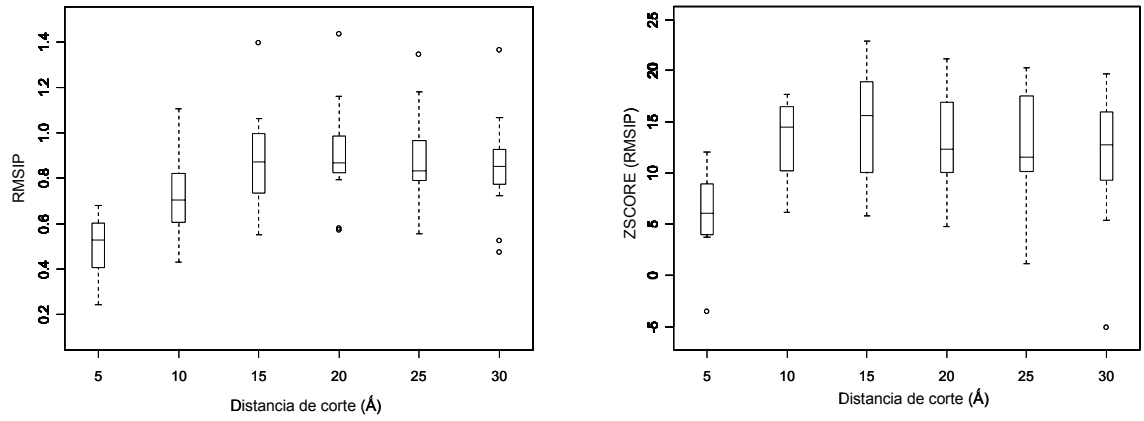


Figura 27. Diagramas de cajas (*box-plots*) para el solapamiento del espacio PCA y el ANM en función de la distancia de corte empleada en el cálculo del ANM. La longitud de las cajas es 1.5 veces el rango intercuartil, dejando fuera los puntos que se van de rango (outliers). (A) Valores del RMSIP. (B) Valores de *Z-score* del RMSIP.

Finalmente, también se estudió si las fluctuaciones por residuo observadas en el *centro estructural* estaban relacionadas con las predichas por los modos normales; es decir, si las regiones que presentan mayores fluctuaciones evolutivas de estructura, corresponden a aquellas determinadas por el ANM como las que experimentan las mayores fluctuaciones vibracionales. Los resultados se pueden ver en la Figura 29.

Para la mayoría de las superfamilias se observó un grado moderado de correlación entre la fluctuación cuadrática media observada en el *centro estructural*, calculada de los alineamientos, y las fluctuaciones predichas por ANM, con coeficientes de correlación de Spearman en el rango de 0.3-0.8 (Figura 29A). Un ejemplo de esta correspondencia se puede ver en la Figura 30. Como se puede apreciar, los perfiles dados por PCA y ANM resultaron similares, aunque con diferentes escalas.

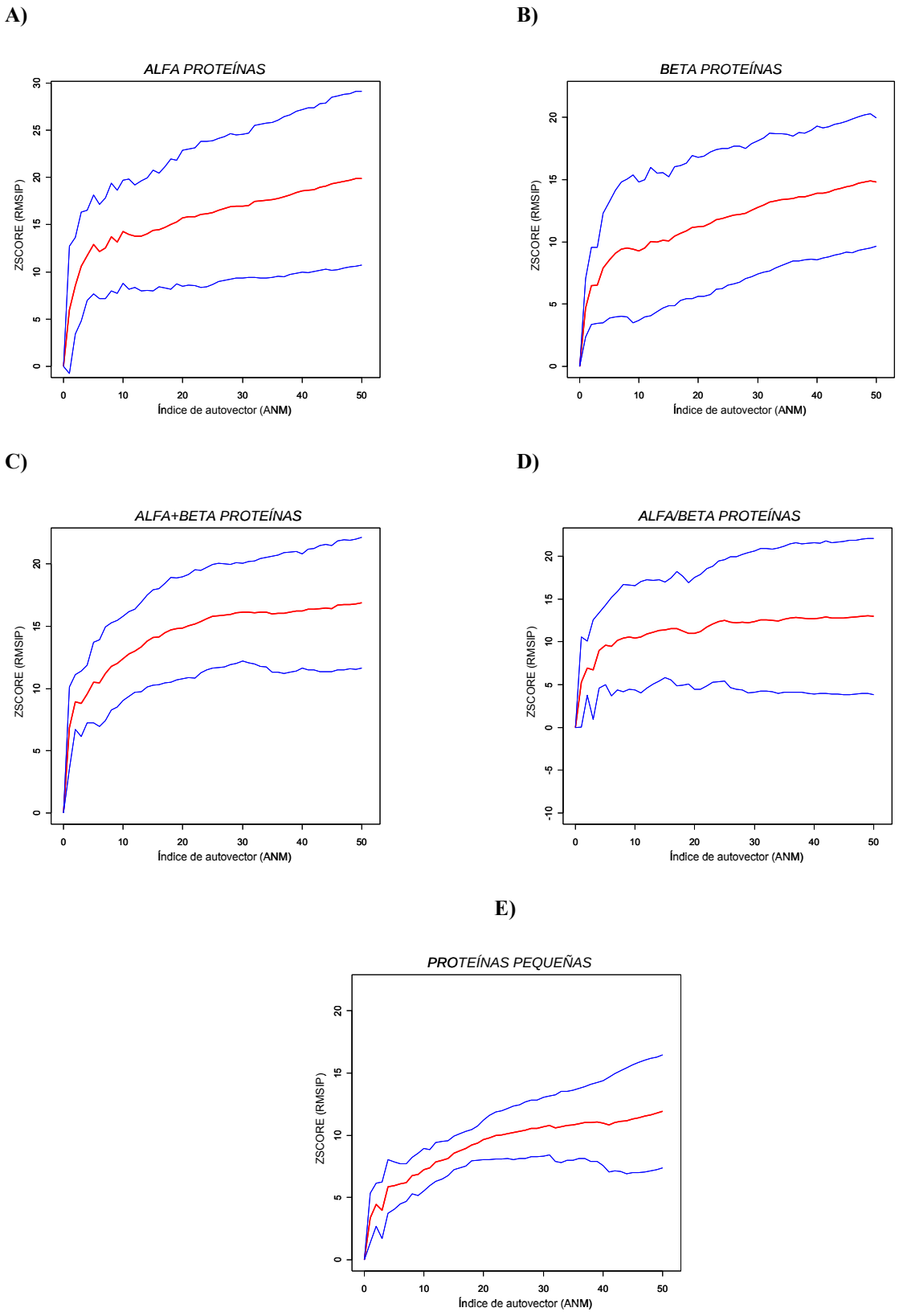


Figura 28. *Z-score* del RMSIP (solapamiento de los espacios PCA y ANM; ver Eq. 28) a la distancia de corte óptima (15 Å), en función del número de modos normales empleados. Sólo se consideraron los 50 modos de más baja frecuencia. (A) Alfa-proteínas; (B) Beta-proteínas; (C) Alfa+Beta proteínas; (D) Alfa/Beta proteínas; y (E) Proteínas pequeñas.

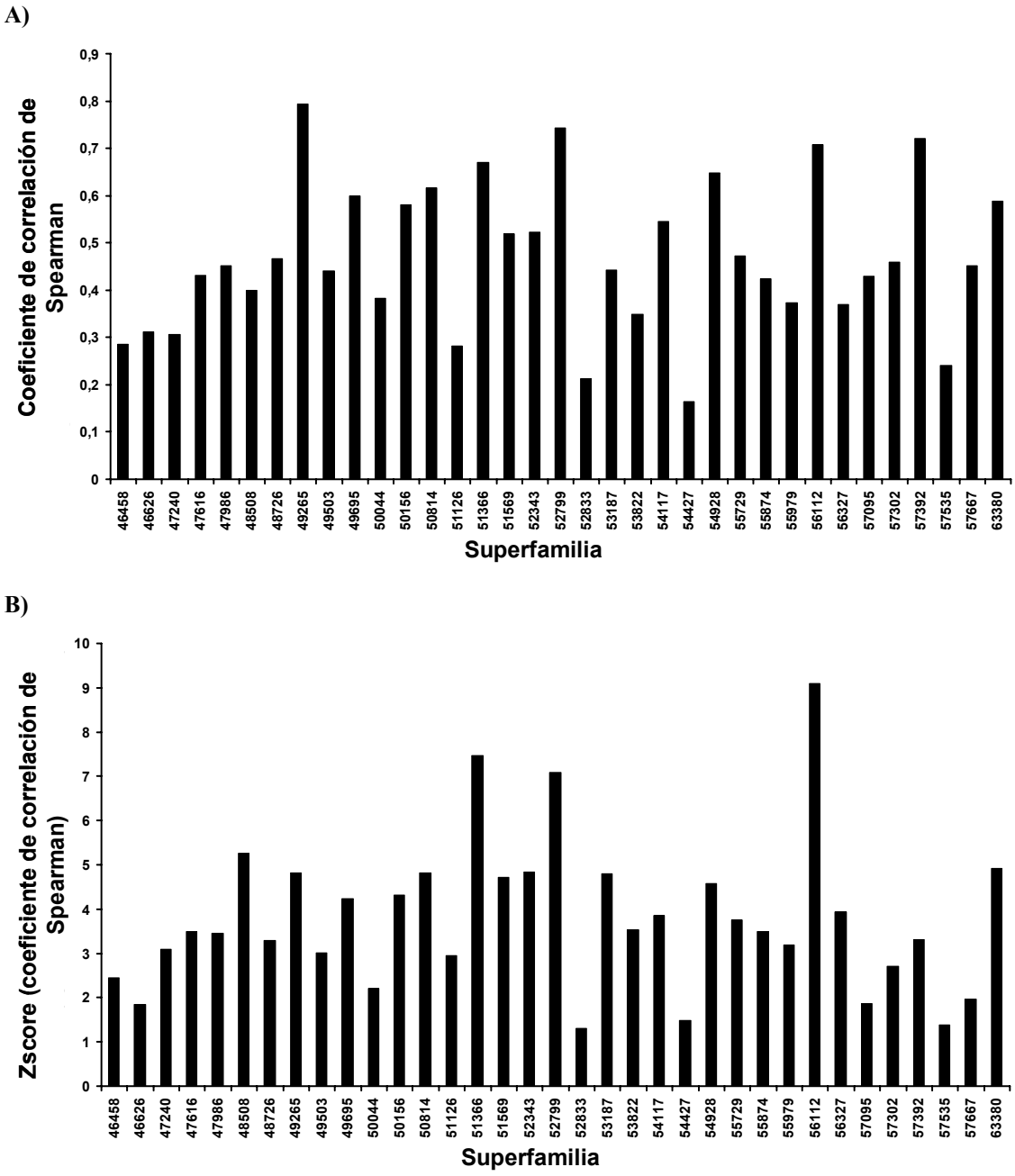


Figura 29. (A) Coeficiente de correlación de rangos de Spearman entre las fluctuaciones cuadráticas medias observadas (a partir de los alineamientos estructurales múltiples) y las calculadas mediante ANM para cada una de las superfamilias estudiadas. (B) El correspondiente *Z-score* de los coeficientes de correlación de rangos de Spearman.

En general, las correlaciones de Spearman encontradas resultaron estadísticamente significativas para todas las superfamilias estudiadas (Figura 29B). Las excepciones fueron citocromo c (46626), NTF2 (54427), thioredoxina (52833), dominio SCR (57535), toxina de escorpión (57095), y dedos de zinc (57667), todos con *Z-scores* menores de 2. En algunos casos, se pudieron encontrar explicaciones para estas desviaciones. Por ejemplo, para el caso del

citocromo c la razón podría ser que el grupo hemo no está incluído en el cálculo de los modos normales de ANM. Para los dominos SCR y toxina de escorpión la explicación posible para la baja correlación podría ser la riqueza en puentes disulfuro que presentan (no considerados en el ANM), y para los dedos de zinc, la presencia de un átomo de Zn que ayuda a mantener su estructura quelando los residuos cisteína e histidina.

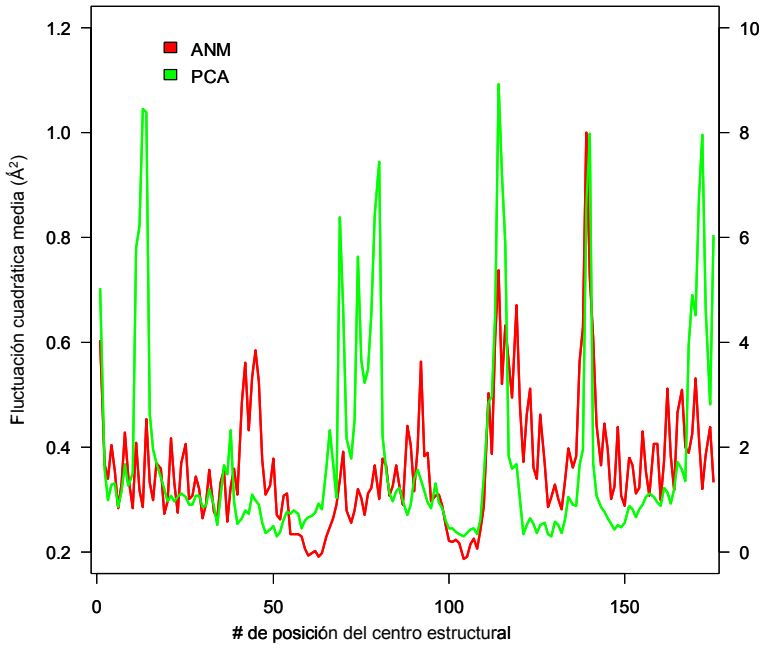


Figura 30. Fluctuación cuadrática media por residuo de *centro estructural* correspondiente a la superfamilia 48508 (dominio del receptor nuclear de unión a ligando). En rojo se representan las fluctuaciones correspondientes al análisis de modos normales, ANM (escala de la izquierda), y en verde al análisis de componentes principales evolutivos, PCA (escala de la derecha).

4.2.3. Comparación EM-PCA y PCA estándar

Se realizó un estudio comparativo entre los espacios PCA y EM-PCA para determinar la ganancia de información que se obtiene al pasar del "*centro estructural estricto*" al "*centro permisivo*" detectados por MAMMOTH-mult. Para ello se estudió el tamaño de la proteína que se puede modelar directamente con EM-PCA en comparación con el que se puede modelar con PCA estándar. Como era de esperar (y así se observó en la Figura 31), los tamaños del *centro estructural permisivo*, que se manejaban cuando se aplicó EM-PCA, fueron mucho mayores que los *centros estrictos* manejados en PCA. Así, se observó que mientras que la moda del tamaño de *centro* era del 80% para el EM-PCA, ésta sólo llegaba al 60% con PCA. De manera similar, se vio que sólo el 22.5% de las proteínas presentó *centros estrictos* mayores del 60%, mientras que en el caso de los *centros* con huecos, este porcentaje llegó hasta el 69%. Por tanto, se demostró que el EM-PCA puede extender considerablemente el alcance del PCA estándar.

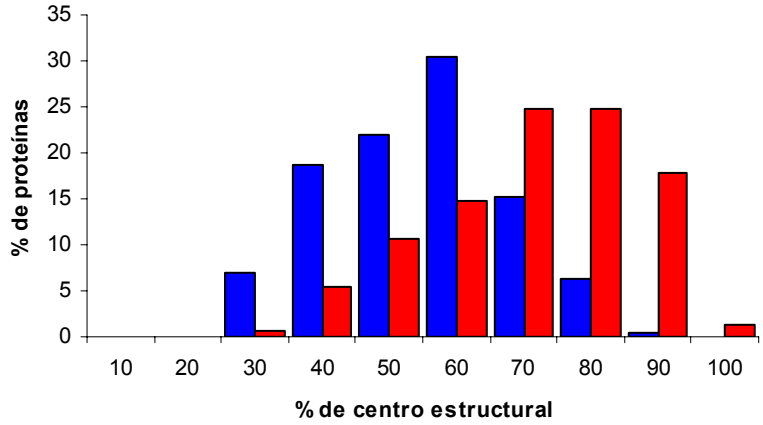


Figura 31. Distribución del tamaño del *centro estructural* para las estructuras correspondientes al conjunto de datos 3.1.3. Azul: *Centro estricto*, modelado con PCA estándar; rojo: *centro permisivo*, modelado con EM-PCA.

4.2.4. Calidad del espacio de muestreo

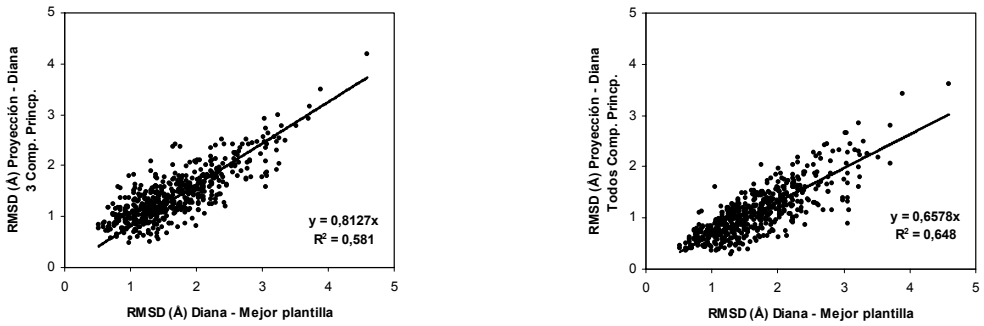
Una vez determinada la extensión de proteína que se puede llegar a modelar utilizando el *centro permisivo* dado por MAMMOTH-mult y EM-PCA, era necesario estudiar con qué calidad se podían llegar a representar las proteínas en este espacio y hasta qué punto, se mejoraban los resultados con la información extra aportada por el espacio ANM. Para ello, se definieron diferentes espacios de muestreo, construidos con distintos conjuntos de autovectores de EM-PCA y de ANM, se proyectaron las estructuras dianas en estos espacios y se midió el RMSD entre las proyecciones obtenidas y las estructuras nativas. La **proyección** es la estructura más parecida a la nativa que se puede encontrar en el espacio de muestreo, y puede considerarse como una medida de la calidad de dicho espacio.

En la Figura 32, se muestran los valores de RMSD para las proyecciones frente a los valores de RMSD entre la estructura nativa y aquella proteína dentro de la familia de homólogos con mayor identidad en secuencia con la diana (considerada como *mejor plantilla*), para todas las proteínas del conjunto de datos 3.1.3.1 (ver apartado de Métodos, pág. 28). De esta manera, se tiene una comparación directa entre la calidad (en términos de RMSD), del mejor modelo que se puede obtener para cada proteína problema mediante los protocolos de modelado por homología estándares (ya que como se ha explicado en otro apartado, el modelo conseguido con estos protocolos casi nunca está más cerca de la proteína diana que la plantilla usada para construirlo), y la del mejor modelo que se puede construir (proyección), usando el espacio de muestreo propuesto en esta tesis.

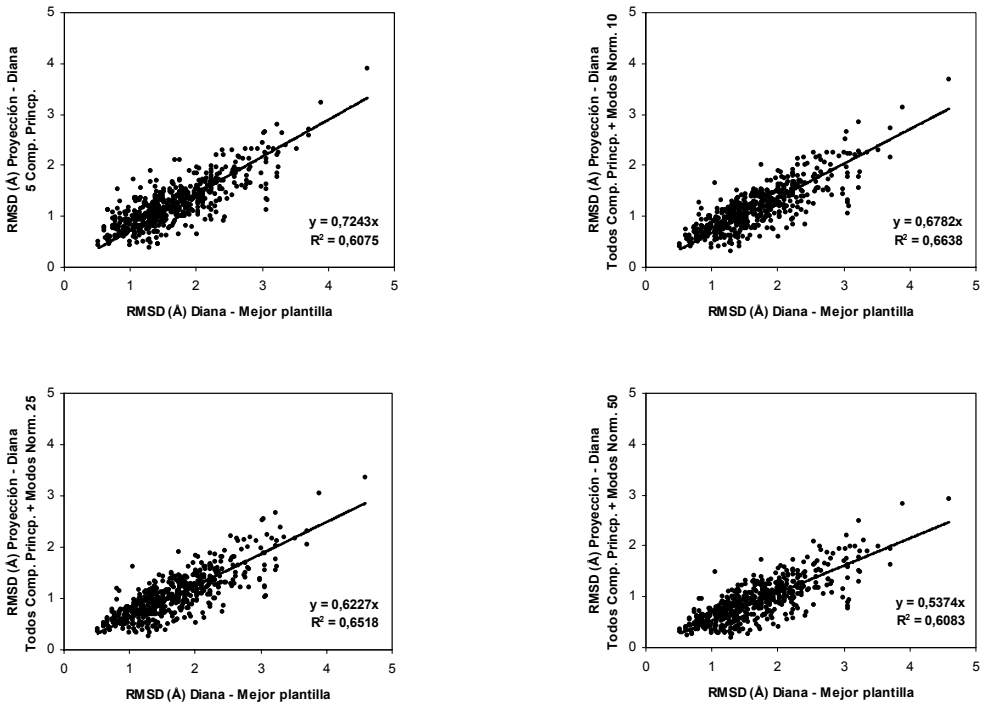
La primera observación es que aunque en el trabajo previo de Qian et al. (Qian et al., 2004) se mostró que las 3 primeras dimensiones del espacio PCA son buenas candidatas para definir un espacio conformacional de dimensionalidad reducida para estudiar las deformaciones

del *centro estructural* de proteínas homólogas, se encuentra que esta aproximación ignora mucha información, puesto que no tiene en cuenta la contenida en el resto de las dimensiones del PCA. Se demostró que el uso de todas las dimensiones EM-PCA proporciona una rebaja considerable de RMSD, ya que para todas las proteínas de este estudio, el valor de RMSD medio del *centro permisivo* entre la estructura de la proyección de la diana en el espacio de tres componentes principales (3-PCs) y la estructura nativa correspondiente, es de aproximadamente 1.40 Å, mientras que el mismo parámetro para el espacio formado por todos los componentes principales es de ~1.09 Å. Por otro lado, se investigó además si una combinación de vectores EM-PCA y ANM podría mejorar estos resultados. Como se describe en el apartado 3.5 de esta tesis (Leo-Macías et al., 2005), se encontró que los espacios de PCA y de ANM de baja frecuencia (hasta 50 modos vibracionales), solapan de manera significativa, por lo que el espacio de ANM podría servir como complemento del espacio de componentes principales evolutivos para explicar información evolutiva desconocida, no presente en los alineamientos múltiples de estructuras debido a un pobre muestreo estructural. En la Figura 32A, se muestran espacios formados por sólo 3 componentes principales (izq.), y por todos los componentes principales (dcha.); en la Figura 32B, el espacio lo componen todos los PC´s más los modos ANM hasta 5, 10, 25 y 50 dimensiones; y la Figura 32C presenta espacios compuestos de sólo 5 y 50 modos de ANM. En todos los casos se obtuvieron pendientes < 1 en las rectas de ajuste, lo que indica que las proyecciones de las proteínas diana en todos los espacios siempre fueron mejores que las estructuras obtenidas eligiendo la *mejor plantilla* según el procedimiento estándar de modelado por homología (como explicado arriba, el de mayor identidad en secuencia con la diana), ya que los RMSD´s de estas últimas siempre resultaron mayores que los de las proyecciones. Asimismo, también se pudo observar que los autovectores de EM-PCA son más eficientes para explicar la varianza evolutiva que los de ANM, puesto que para obtener estructuras similares a la nativa de la misma precisión, se necesitaron más vectores vibracionales que evolutivos. Esto coincide con nuestras observaciones previas (Leo-Macías et al., 2005).

A)



B)



C)

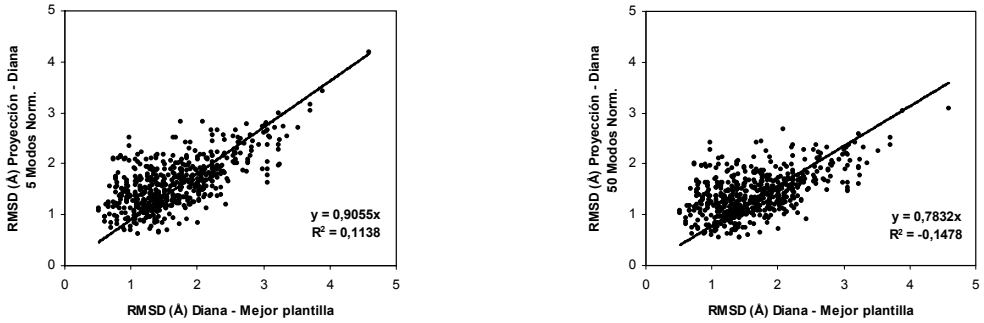


Figura 32. Correlaciones encontradas para la región del *centro permisivo* entre el RMSD del homólogo más cercano y el RMSD de la proyección. (A) Resultados EM-PCA usando tres (izquierda), o todas las dimensiones disponibles (derecha); (B) Resultados EM-PCA-ANM usando 5, 10, 25 y 50 dimensiones; C) Resultados utilizando sólo vectores del espacio ANM, con 5 y 50 dimensiones.

Del análisis de la Figura 32, se desprendió que una buena opción de espacio de muestreo, consiste en un espacio mixto formado por todos los vectores de EM-PCA más los vectores de ANM necesarios hasta completar en total unas 50 dimensiones (espacio EPA), ya que fue el caso para el que se obtuvo la mejor pendiente en la recta de ajuste. En este espacio, el RMSD promedio entre las proyecciones y las estructuras nativas fue de ~ 0.87 Å.

Para resumir estos hallazgos, en la Figura 33, se comparó la calidad del espacio utilizado por Qian et al. (3-PC´s), con el espacio EPA. La mejora de calidad al pasar de uno a otro es considerable, ya que, como se pudo observar, cuando sólo se utilizan 3 componentes principales, sólo un 18.5% aproximadamente de las proteínas tienen estructuras parecidas a la nativa con un RMSD < 1 Å, mientras que en el espacio EPA, el porcentaje para el mismo nivel de precisión es de un 65.6%; y éste sube a 92.3% para RMSD < 1.5 Å.

Por tanto, el espacio EPA parece ser una buena elección como subespacio de muestreo a nivel de cadena principal no sólo en cuanto a su calidad (puede representar la mayoría de las proteínas dentro de 1 Å de diferencia en RMSD), sino también en cuanto a su dimensionalidad (baja, tan solo 50 dimensiones).

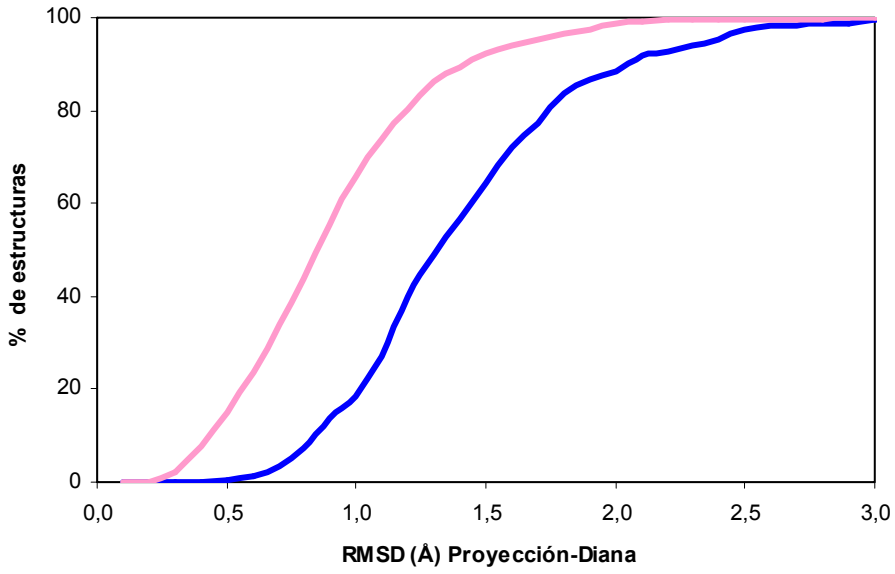


Figura 33. RMSD del *centro permisivo* frente a la fracción acumulada de dianas por debajo de ese valor de RMSD de corte, usando EM-PCA con tres componentes (línea azul), o usando el espacio EM-PCA-ANM con 50 dimensiones (línea rosa).

4.2.5. Dianas de CASP5

Una evaluación adicional del espacio EPA obtenido, consistió en determinar su comportamiento utilizando las mismas proteínas diana utilizadas en CASP5. Se utilizó exactamente el mismo conjunto de 67 dianas divididas en dominios por los organizadores de la competición, a excepción de la diana *T0186_3*, para la que la búsqueda con MAMMOTH de

pares en la base de datos de estructuras construida al efecto (ver apartado de Métodos, pág. 28), resultó infructuosa y por tanto, no pudo ser modelada.

Este mismo conjunto de proteínas fue utilizado por (Contreras-Moreira et al., 2005) en su evaluación de los límites empíricos de la metodología del modelado por homología, lo que nos permitió comparar a su vez, el comportamiento de nuestro espacio EPA con respecto a dichos límites.

En la Tabla 4, Tabla 5 y Tabla 6 se muestra el resumen de los resultados de la construcción de los espacios de muestreo para las dianas, divididos en tres categorías según la "*dificultad de modelado*" de la diana correspondiente. Para determinar esta dificultad se usó el promedio de la puntuación de MAMMOTH de pares de los alineamientos estructurales de cada diana con cada una de sus plantillas encontrada en la base de datos (*Ave_mam_p*) (ver apartado de Métodos, pág. 28). Así, se consideraron "*dianas fáciles*" aquellas para las que $Ave\_mam\_p \geq 12$ (Tabla 4); "*dianas moderadas*", si $7 \leq Ave\_mam\_p < 12$ (Tabla 5), y "*dianas difíciles*", si $4 \leq Ave\_mam\_p < 7$ (Tabla 6). Por debajo de 4, la similitud estructural no es significativa y por tanto, no se seleccionaron plantillas en este régimen.

Diana	*# Plantillas*	*# Tot res*	*# res centro*	*# res lazos*	*# Mod res*	*% Mod str*
T0179_2	24	218	82	9	91	41,74
T0138	15	135	65	19	84	62,22
T0149_1	4	201	104	25	129	64,18
T0150	3	96	90	5	95	98,96
T0191_2	10	143	85	12	97	67,83
T0169	10	156	94	23	117	75,00
T0142	6	280	87	29	116	41,43
T0154_1	5	185	118	14	132	71,35
T0136_2	7	264	144	14	158	59,85
T0172_1	9	192	106	15	121	63,02
T0136_1	6	256	145	15	160	62,50
T0153	4	134	113	8	121	90,30
T0167	4	180	105	28	133	73,89
T0155	3	117	105	3	108	92,31
T0137	19	133	115	11	126	94,74
T0183	28	247	116	17	133	53,85
T0165	15	318	146	21	167	52,52
T0185_2	5	197	124	16	140	71,07
T0178	32	219	139	21	160	73,06
T0189	7	319	95	35	130	40,75
T0182	6	249	148	16	164	65,86

Tabla 4. Resumen de los resultados para las "*dianas fáciles*" de CASP5. *#Plantillas* es el número de plantillas encontradas en la base de datos; *#Tot res* es el número total de residuos de la diana; *#res centro* es el número de residuos en el *centro*; *#res lazos* es el número de residuos en *lazos* que fueron modelados; *#Mod res* es el número total de residuos modelados; *%Mod str* es el porcentaje de estructura de la diana que corresponde a ese número de residuos modelado.

Diana	*# Plantillas*	*# Tot res*	*# res centro*	*# res lazos*	*# Mod res*	*% Mod str*
T0146_2	6	89	61	18	79	88,76
T0162_1	3	56	28	11	39	69,64
T0187_2	3	227	59	10	69	30,40
T0132	3	147	53	12	65	44,22
T0185_3	3	130	44	15	59	45,38
T0187_1	5	187	72	6	78	41,71
T0184_2	5	72	47	13	60	83,33
T0188	4	107	62	7	69	64,49
T0159_1	3	167	70	22	92	55,09
T0130	4	100	61	9	70	70,00
T0148_1	6	71	58	10	68	95,77
T0157	7	120	44	9	53	44,17
T0148_2	4	91	52	19	71	78,02
T0177_3	3	75	54	9	63	84,00
T0162_3	3	168	66	12	78	46,43
T0174_2	3	155	103	19	122	78,71
T0143_2	47	95	70	17	87	91,58
T0173	4	287	92	10	102	35,54
T0151	9	106	74	11	85	80,19
T0185_1	7	101	68	13	81	80,20
T0191_1	6	139	54	12	66	47,48
T0154_2	5	103	73	23	96	93,20
T0143_1	51	121	91	15	106	87,60
T0184_1	3	165	33	1	34	20,61
T0193_2	16	130	57	10	67	51,54
T0147	25	234	82	24	106	45,30
T0186_2	12	250	109	22	131	52,40

Tabla 5. Resumen de los resultados para las "*dianas de dificultad moderada*" de CASP5.

Diana	*# Plantillas*	*# Tot res*	*# res centro*	*# res lazos*	*# Mod res*	*% Mod str*
T0141	3	187	44	6	50	26,74
T0179_1	3	56	30	6	36	64,29
T0172_2	3	101	54	7	61	60,40
T0181	9	111	53	11	64	57,66
T0162_2	4	51	27	8	35	68,63
T0186_1	5	77	36	6	42	54,55
T0156	3	156	51	14	65	41,67
T0168_1	3	170	76	23	99	58,24
T0174_1	4	197	67	16	83	42,13
T0193_1	5	74	36	5	41	55,41
T0168_2	4	141	55	9	64	45,39
T0161	3	154	45	3	48	31,17
T0146_1	5	107	44	15	59	55,14
T0177_2	29	88	39	6	45	51,14
T0177_1	8	57	36	2	38	66,67
T0170	8	69	49	6	55	79,71
T0176	4	100	41	9	50	50,00
T0159_2	3	142	71	19	90	63,38
T0149_2	4	116	38	9	47	40,52

Tabla 6. Resumen de los resultados para las "*dianas difíciles*" de CASP5.

Como es de esperar, en general, se encontraron más plantillas para las *dianas fáciles* que para las *difíciles*, aunque en todos los casos el número encontrado fue suficiente (al menos 3) y permitió obtener *centros estructurales* suficientemente grandes, que abarcaban más del 60% de la estructura para más de la mitad de los casos (Figura 34). El número de residuos en los *centros* junto con aquellos que se encontraban en *lazos* que fueron modelados con MODELLER (ver apartado de Métodos, pág. 47), permitieron generar modelos para regiones suficientemente grandes de las dianas (en torno al 62% en promedio). Tan solo hay dos casos, el *T0141* y el *T0184_1* para los que no se llegó al 30% de estructura modelada, debido a la dificultad de encontrar plantillas adecuadas.

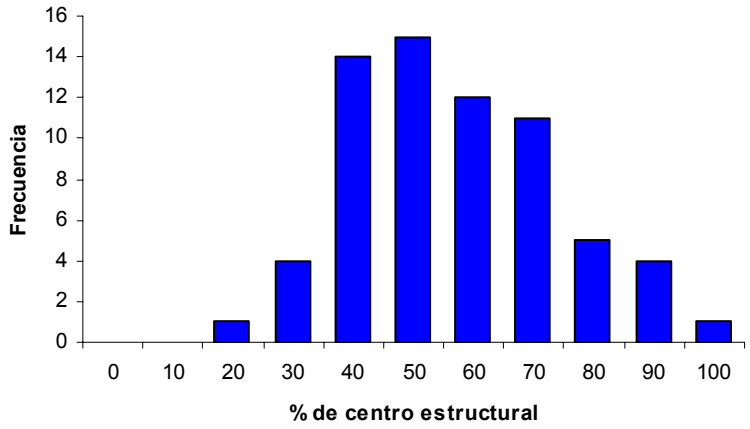


Figura 34. Distribución del tamaño de *centro permisivo* encontrado para las dianas de CASP5 estudiadas.

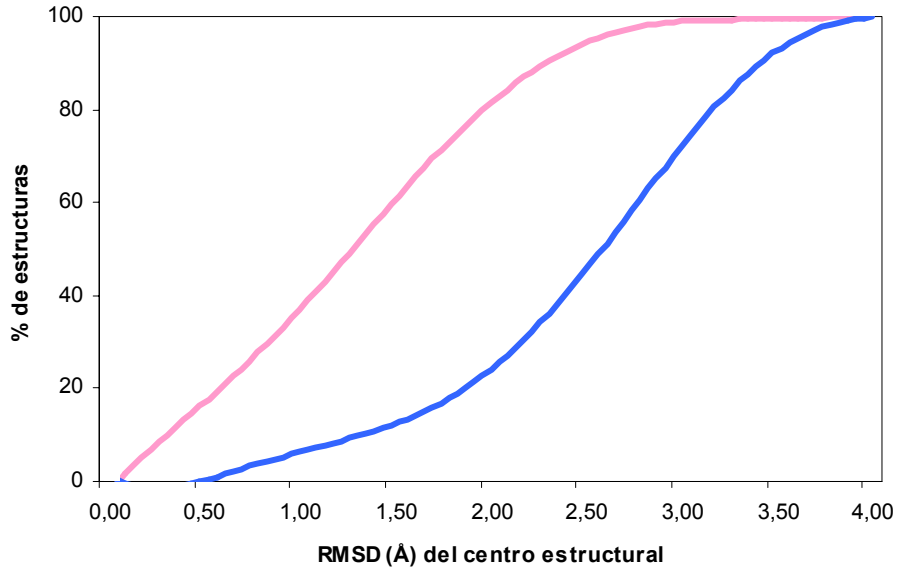


Figura 35. RMSD del *centro permisivo* frente a la fracción acumulada de dianas por debajo de ese valor de RMSD de corte para las 67 dianas de CASP5. Línea azul: RMSD *diana-mejor plantilla*; línea rosa:. *diana-modelo* construido a partir de la proyección.

Al proyectar cada diana en su correspondiente espacio EPA y medir el RMSD del *centro estructural* de dicha proyección con respecto a la estructura nativa, se observó que aproximadamente el 35% de las dianas se pudo representar en su espacio EPA con un RMSD < 1 Å, y el 60% de las mismas para un RMSD < 1.5 Å (Figura 35). Aunque estos valores son menores que para el conjunto de 547 proteínas estudiado previamente (Figura 33), debido a la especial dificultad de modelado de algunas dianas de esta competición, los porcentajes siguen siendo mayores que si se comparan con los valores para las mejores plantillas posibles, en donde el porcentaje de casos con un RMSD < 1.5 Å no llegó al 20%.

Por otra parte, para determinar la calidad de los modelos obtenidos a partir de la proyección para cada una de las dianas, se utilizó también el parámetro GDT_TS (ver Métodos, pág. 48). Este parámetro permitió comparar los resultados obtenidos mediante el protocolo desarrollado en esta tesis con los de los grupos que realizaron la mejor actuación en CASP5, así como con los mejores modelos construidos a partir de fragmentos utilizando FRAGBENCH (Contreras-Moreira et al., 2005), que proporcionan una idea de los límites intrínsecos de la metodología de modelado a partir de patrones. En la Figura 36 se presentan los resultados de esta comparativa. Se observó que por debajo del valor de corte del 40% de identidad en secuencia, el modelado a partir de la proyección en el espacio EPA alcanzó los límites empíricos de la metodología, ya que sus resultados y los de FRAGBENCH presentaron curvas de ajuste muy similares y dieron las mismas mejoras en los valores de GDT_TS con respecto a los protocolos actuales usados por los grupos que mejores resultados obtuvieron en CASP5. Esto resultó muy interesante, ya que es precisamente en esta región de baja identidad en secuencia donde se dan las mayores dificultades en esta metodología, donde los modelos construidos suelen tener menos de la mitad de sus Cα a una distancia menor de 3.5 Å con respecto a sus posiciones correctas. Por encima del 40% de identidad en secuencia, los métodos actuales generan modelos de una calidad comparable a la de las estructuras experimentales de baja resolución y se apreció que el uso del espacio EPA no produjo ninguna mejora significativa con respecto a ellos.

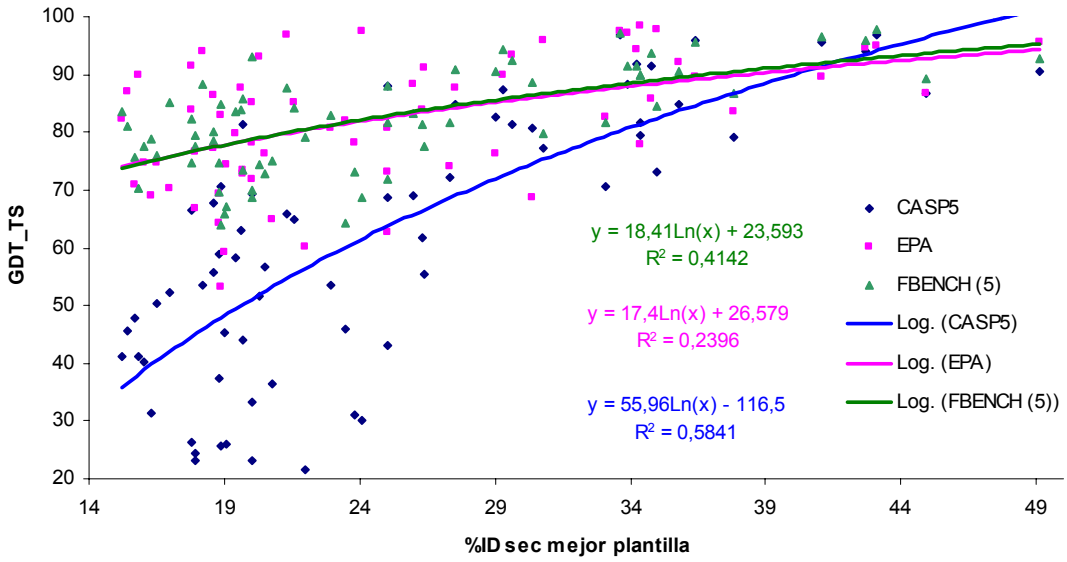


Figura 36. Comparativa para la calidad de los modelos para las 67 dianas de CASP5 estudiadas. Resultados de GDT_TS para los modelos obtenidos a partir de la proyección de las dianas en el espacio EPA construido para cada uno de ellos (en rosa), para los mejores modelos enviados a CASP5 por los diferentes grupos participantes (en azul), y para las mejores soluciones basadas en patrones producidas por FRAGBENCH usando fragmentos de tamaño 5 (en verde). Se representa el valor de GDT_TS frente al porcentaje de identidad en secuencia con la *mejor plantilla* (en términos de secuencia).

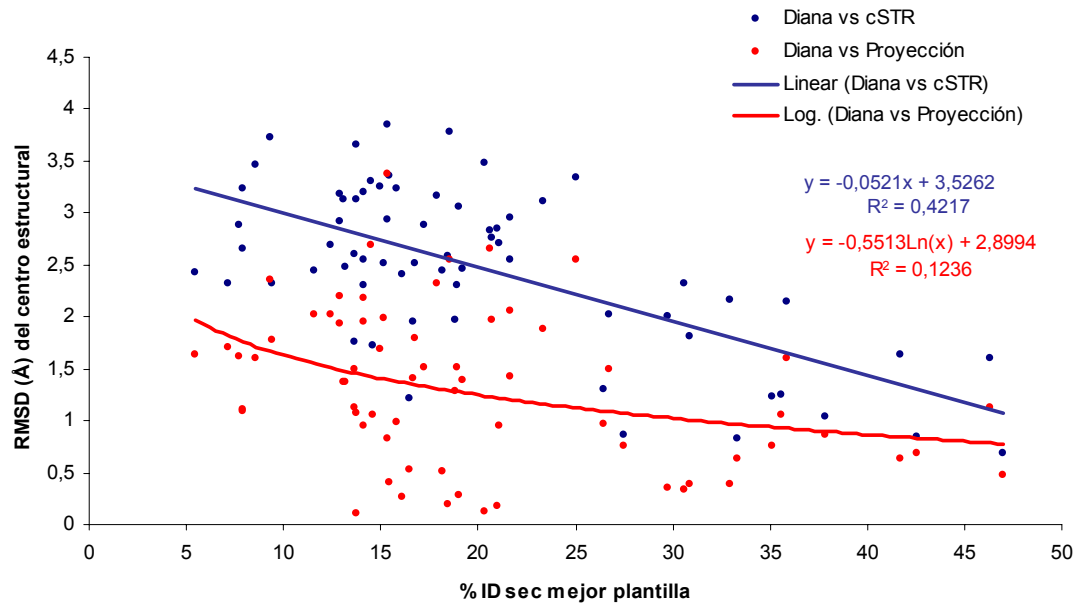


Figura 37. Valores de RMSD para los *centros estructurales* de los modelos construidos a partir del espacio EPA y las dianas originales, comparado con los valores obtenidos para la misma región, entre la diana y la *mejor plantilla* posible (en términos de similitud estructural), representado en función del porcentaje de identidad en secuencia entre ambos.

Asimismo, se observó (Figura 37) que el espacio EPA permitió obtener mejores *centros estructurales* en términos de RMSD que la mejor plantilla posible en todos los casos, especialmente en la región entre el 10-40% de identidad en secuencia, donde se apreciaron las mayores diferencias. Esto volvió a poner de manifiesto la utilidad de este espacio especialmente en las regiones de baja identidad en secuencia, en donde la *mejor plantilla* y la diana tienen estructuras suficientemente diferentes como para que la información estructural extra aportada por el resto de las plantillas a partir de las cuales se construye el espacio EPA, resulte muy beneficiosa en la construcción del modelo final. Sin embargo, en las regiones de alta identidad en secuencia, la similitud estructural entre la diana y la plantilla suele ser lo suficientemente grande como para que el aporte de información estructural extra no resulte tan crucial en la mejora de la calidad de los modelos.

Hasta aquí, sólo se ha tenido en cuenta la traza de Cα de la proyección de cada diana en su espacio EPA correspondiente. Dados los resultados prometedores obtenidos, se abordó también el estudio de la calidad de los modelos finales construidos de la forma más completa posible, que se puden obtener a partir de esta proyección de la traza de Cα. Para ello, se añadieron los átomos necesarios para completar la cadena principal y las cadenas laterales y además del RMSD, se utilizaron también parámetros del mapa de Ramachandran, ángulos chi-1 y chi-2 y número de malos contactos para evaluar la calidad de los modelos finales completos, los cuales fueron sometidos también a un proceso extra de minimización con AMBER para estudiar su efecto (ver apartado de Métodos, pág. 48).

En la Tabla-Mat.Sup. 2 se detallan los resultados de la evaluación y en la Figura 38 se presenta un resumen de los mismos. En general, se apreció un aumento del RMSD tanto del *centro estructural* (Figura 38B) como del modelo completo (Figura 38A), cuando éste se somete a la minimización, efecto que no se observó en el caso de los residuos en *lazos* (Figura 38C). El número de malos contactos entre átomos cayó prácticamente a 0 (Tabla-Mat.Sup. 2) y el número de residuos en las zonas más favorecidas del mapa de Ramachandran aumentó después de la minimización (Figura 38D). El valor de los ángulos chi pareció no verse afectado (Figura 38E y Figura 38F).

A la vista de la Tabla-Mat.Sup. 2, la Tabla 4, Tabla 5 y Tabla 6 se puede concluir que el espacio EPA permitió la obtención de resultados muy satisfactorios en los tres rangos de dificultad de modelado en que se dividieron las dianas, ya que en los tres casos se pudieron obtener modelos con un elevado porcentaje de estructura modelada y un RMSD del *centro estructural* pequeño. Ejemplos de estos casos son la diana *T0150* para la zona "*fácil*" (99% de estructura modelada y $RMSD_{centro}$ = 0.39 Å); la diana *T0184_2* para la zona "*moderada*" (83.3% de estructura modelada y $RMSD_{centro}$ = 0.33 Å) y la diana *T0170* para la zona "*difícil*" (79.8% de estructura modelada y $RMSD_{centro}$ = 0.11 Å). En la Figura 39 se pueden observar las superposiciones estructurales de estos ejemplos.

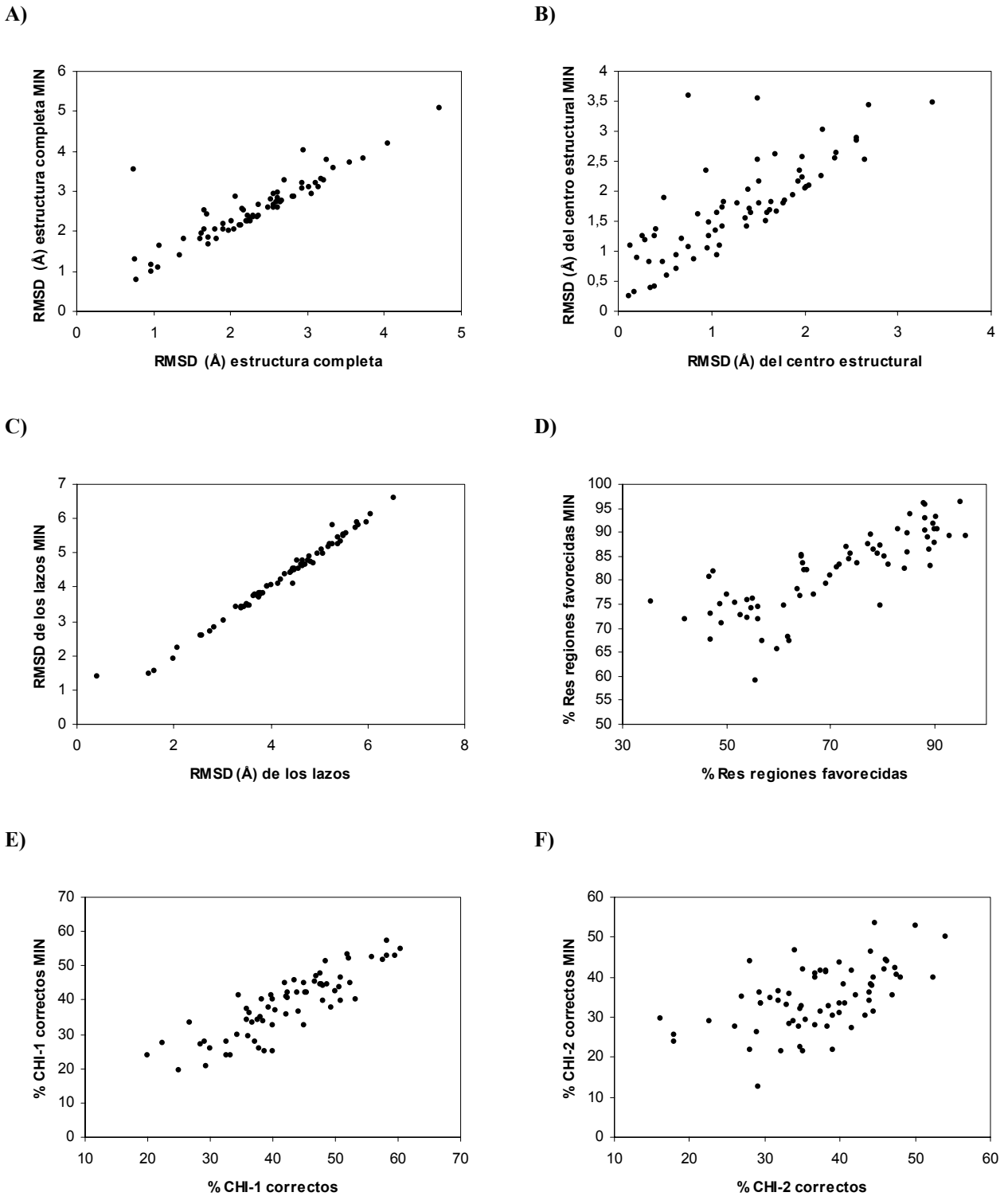
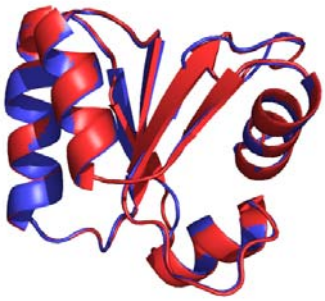


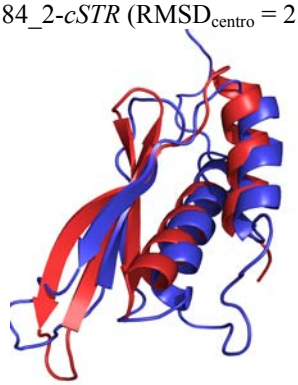
Figura 38. Resumen de los resultados de la reconstrucción de los modelos completos antes y después de la minimización. Se representa el RMSD para los Cα entre los modelos y las estructuras nativas A) para toda la estructura modelada, B) para la región de *centro estructural*, C) para los *lazos*; D) porcentaje de residuos en las zonas permitidas del mapa de Ramachandran, E) % de ángulos chi-1 correctos y F) chi-2 correctos.

A) T0150-*cSTR* ($RMSD_{centro}$ = 2.16 Å)

B) T0150- Modelo ($RMSD_{centro}$ = 0.39 Å)



C) T0184_2-*cSTR* ($RMSD_{centro}$ = 2.33 Å)



D) T0184_2-Modelo ($RMSD_{centro}$ = 0.33 Å)

E) T0170-*cSTR* ($RMSD_{centro}$ = 3.65 Å)

F) T0170-Modelo ($RMSD_{centro}$ = 0.11 Å)

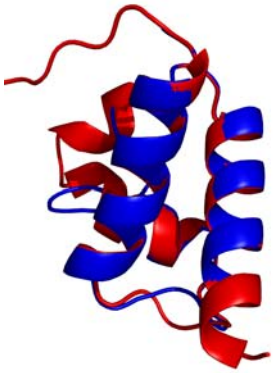


Figura 39. Ejemplos de modelos de dianas de CASP5 superpuestos con MAMMOTH de pares con su estructura nativa correspondiente. En rojo se muestra la *diana* y en azul tanto la *mejor plantilla* en términos estructurales como el *modelo final* construido a partir de la proyección.

Se muestran varios ejemplos de la calidad de los modelos completos reconstruidos para las dianas, en comparación con la *mejores plantillas* disponibles en términos estructurales (*cSTR*), en el momento en que tuvo lugar CASP5. Como se puede apreciar, la mejora de RMSD fue considerable para el *centro estructural* de los modelos reconstruidos a partir de las proyecciones de las dianas en el espacio EPA. La Figura 40 muestra un ejemplo de la calidad de las cadenas laterales modeladas. Se apreció que cuando los residuos pertenecen a las zonas centrales de los elementos de estructura secundaria del modelo, las cadenas laterales pueden modelarse con una calidad satisfactoria, mientras que para los residuos en los extremos de estos elementos, las cadenas laterales no se modelan bien.

A) **B)**

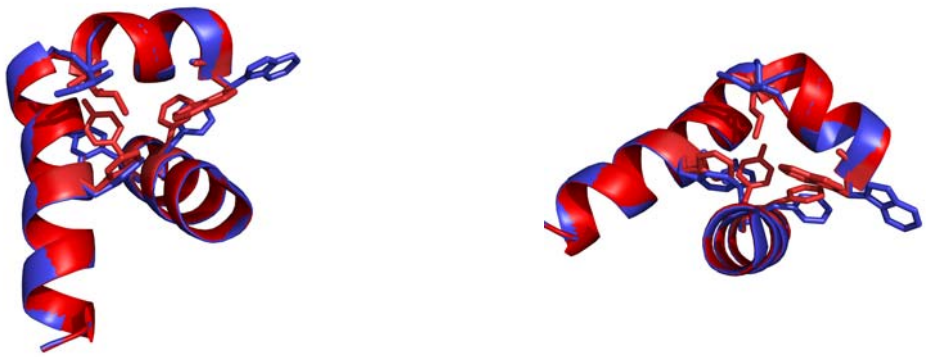


Figura 40. Ejemplos de las cadenas laterales de los residuos hidrofóbicos del *centro estructural* de la diana *T0170* en dos vistas diferentes. La diana se representa en rojo y el modelo reconstruido a partir de la proyección en azul.

4.3. Eficiencia del algoritmo de muestreo

Una vez determinado un subespacio de baja dimensionalidad en el que las estructuras de proteínas se pueden representar con suficiente precisión, resultó necesario el desarrollo de un método de muestreo en este subespacio. Para ello, se estudió la eficiencia en el muestreo de un algoritmo de intercambio de réplicas de Monte Carlo (REMC) (ver Métodos, pág. 60). En la primera evaluación, la función de energía se simplificó de manera que consistiera únicamente en el RMSD entre la estructura muestreada y la problema (diana). Se permiten movimientos tanto del *centro estructural* de la proteína como de los *lazos*, por lo que se pudo modelar la cadena completa. El *centro estructural* de la proteína se movió en el espacio EPA (EM-PCA-ANM), mientras que los *lazos* lo hicieron en un espacio *no*-EPA y se modelaron mediante el algoritmo CCD (ver apartado de Métodos, pág. 55). Para verificar la eficiencia del muestreo, la estructura seleccionada como diana se excluyó del cálculo de los componentes principales. En la Figura 41 se comparó el RMSD de la diana y la proyección (es decir, de la solución analítica de la proteína problema en el espacio EPA), con el de la diana y el modelo obtenido mediante simulación REMC para todas las proteínas del conjunto de datos 3.1.3.1. Al realizar un ajuste lineal, se observó que tanto la pendiente como el cuadrado del coeficiente de correlación entre los dos conjuntos de datos eran muy cercanos a la unidad, lo que indica un buen comportamiento de este algoritmo de muestreo, ya que las estructuras encontradas estaban muy cerca de las proyecciones de las estructuras nativas en el espacio EPA (equivalentes a las mejores soluciones posibles en dicho espacio). En la Figura 42, se representan los resultados del muestreo para toda la cadena de la proteína, incluyendo tanto el *centro estructural* como los *lazos*. Se observó que el 79.3% de todos los modelos tenían un RMSD < 1 Å con respecto a la estructura nativa. Es interesante resaltar que éste es mejor que el valor obtenido cuando sólo se considerabó el *centro estructural* de la proyección (sin *lazos*) (Figura 33), en donde este

porcentaje era de 65.6% (ver Resultados, pág. 89). Ello puede deberse a que el tipo de movimiento de los *lazos* en la simulación está menos restringido que la parte del *centro estructural* y así es posible ajustarlos con mayor libertad a la estructura nativa.

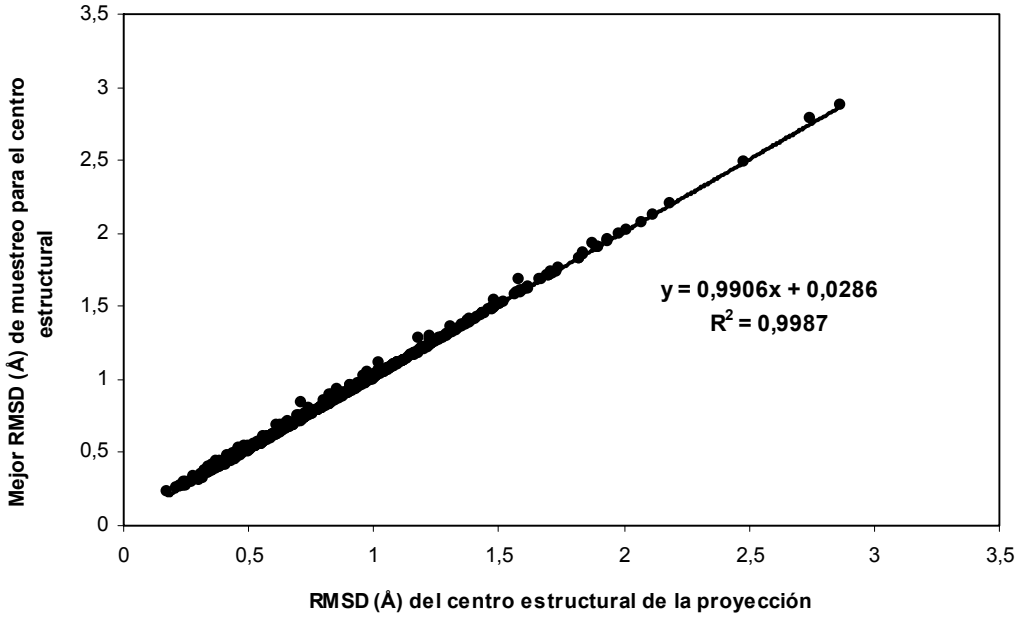


Figura 41. Correlación entre el RMSD del *centro permisivo* entre el modelo obtenido por proyección y el obtenido por simulación REMC para todas las proteínas del conjunto.

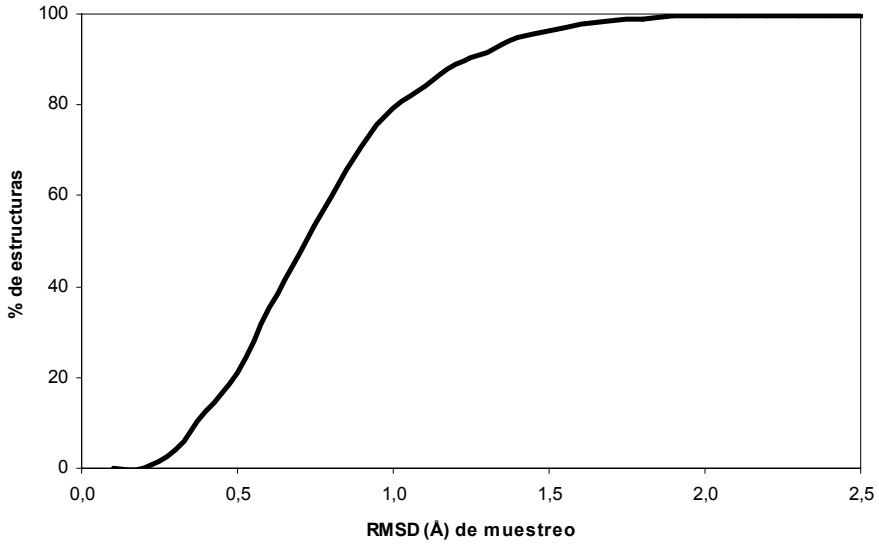


Figura 42. RMSD de la cadena principal entre el modelo y la diana frente a la fracción acumulada de estructuras por debajo de ese valor de corte de RMSD para las simulaciones REMC.

Ambas gráficas (Figura 41 y Figura 42), muestran una gran correlación entre el mínimo global en el espacio, que corresponde a la estructura de la proyección, y la mejor estructura obtenida mediante simulación. Hasta cierto punto, esto indica la validez del espacio de muestreo

definido y de su combinación con el método de REMC, y sirve como demostración de que REMC es capaz de alcanzar el mínimo global en el espacio EPA en todos los casos.

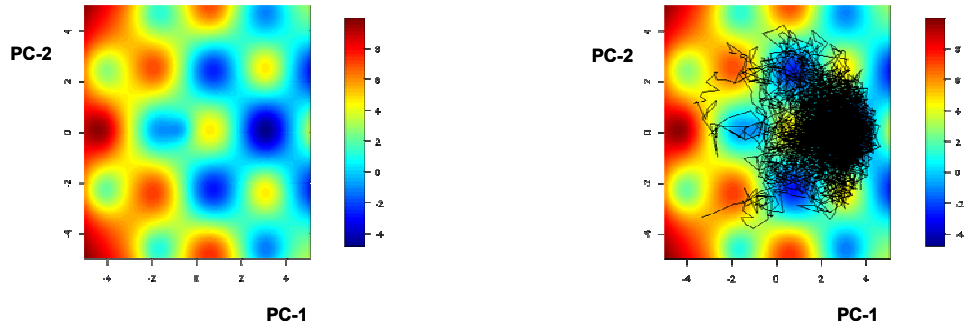
4.3.1. Estudio del comportamiento del protocolo EPA-REMC en condiciones más realistas. Fiabilidad y validez del método

La función de energía utilizada hasta ahora (el RMSD), no es práctica, puesto que en una situación real, no se dispone de la estructura de la diana con la que calcular el RMSD, aunque sirvió para verificar la correcta implementación del algoritmo. Por esta razón, se decidió explorar el comportamiento del protocolo EPA-REMC bajo condiciones más realistas. Para ello, se crearon perfiles energéticos de diferentes grados de dificultad y se midió la capacidad del algoritmo de muestreo para alcanzar el mínimo global (ver apartado de Métodos, pág. 60). La Figura 43 ilustra este procedimiento con un ejemplo.

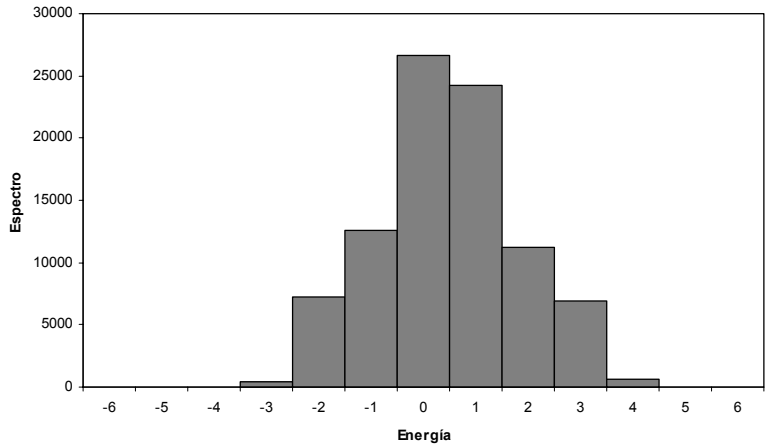
En la Figura 43A se representaron los resultados muestreados en una superficie de energía con un total de 2120 picos de ruido. El valor de salto de energía es *Esalto* = 0.5 con su correspondiente *Zscore*= -3.437. El muestreo se representó en una sección bidimensional del espacio correspondiente a los 2 primeros componentes principales. Esta sección bidimensional representa bien las características de toda la superficie de energía y como puede apreciarse, el muestreo realizado resultó muy eficiente, ya que se acercó rápidamente al mínimo global y a continuación, muestreó la región vecina.

La Figura 43B muestra el espectro energético de la superficie de energía diseñada y en la Figura 43C se presentaron las distribuciones canónicas de probabilidad para las energías muestreadas a 8 temperaturas. Como puede apreciarse, en el muestreo existió el solapamiento necesario entre distribuciones vecinas, que indicó que hubo un número suficiente de intercambio de réplicas, de manera que se pudieron superar las barreras de energía y saltar fuera de los mínimos locales a temperaturas altas y alcanzar los mínimos a temperaturas bajas. Esto evitó la posibilidad de quedarse atrapado en mínimos locales; por tanto, el muestreo realizado fue ergódico y pudo cubrir en mayor o menor medida toda la superficie de energía.

A)



B)



C)

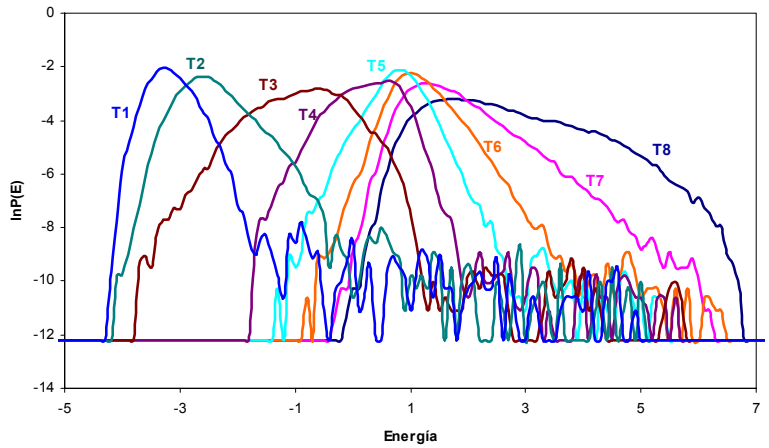


Figura 43. Ejemplo de la evaluación del muestreo REMC. (A) Proyección 2D de la superfecie de energía; en azul se muestran las regiones de mínimos y en rojo la de máximos; (B) espectro energético de la superficie de energía; (C) Distribución de la energía explorada en las simulaciones REMC.

También se comparó el comportamiento del método REMC como algoritmo de muestreo en el espacio EPA con respecto a otros métodos de búsqueda comúnmente usados, RS (*Random Search*, búsqueda aleatoria) y SA (*Simulated Annealing*, templado simulado). Se presenta aquí un ejemplo de los resultados de esta comparativa.

En este ejemplo, se modelaron superficies con dos densidades de ruido diferentes. En una, la distancia entre los centros de ruido cercanos fue de 2.574 Å, resultando en un total de 2120 picos de ruido, y en la otra esta distancia fue de 2.403 Å, dando 2903 picos. Para cada tipo de densidad de ruido se modelaron superficies con 2 tipos de salto de energía, *Esalto* = 0.5 y 1.5. Para cada tipo de superficie se realizan 10 cálculos y se calcula la media y la desviación estándar de los resultados (Tabla 7). Se muestra el correspondiente *Zscore*. Los parámetros de muestreo son los mismos que los usados en el ejemplo mostrado en la Figura 43 (ver también apartado de Métodos, pág. 60). Se muestran los mínimos globales de energía, *Eglobal*, para las diferentes superficies, los cuales varían poco debido a diferentes solapamientos de los picos gaussianos vecinos. Se guardaron dos tipos de resultados: *Emin*, que es la energía más baja de las estructuras muestreadas y *rmsEmin*, que es el RMSD correspondiente entre la estructura muestreada de menor energía y la diana. Los pasos de muestreo para los tres casos, REMC, SA y RS se ajustaron a 10^5 y la temperatura descendió de 2.0 a 0.05 para SA y REMC.

		Nº picos=2120		Nº picos=2903	
	Zscore	-3.437	-3.221	-3.395	-3.267
	Eglob	-4.765	-4.817	-4.778	-4.828
REMC	<Emin>	-4.199 (0.056)	-4.225 (0.063)	-4.183 (0.058)	-4.225 (0.064)
REMC	<rmsEmin>	0.381 (0.021)	0.387 (0.021)	0.366 (0.028)	0.383 (0.027)
SA	<Emin>	-1.621 (0.251)	-1.690 (0.187)	-1.523 (0.416)	-1.758 (0.141)
SA	<rmsEmin>	1.047 (0.077)	1.031 (0.067)	1.032 (0.114)	1.011 (0.042)
RS	<Emin>	0.156 (0.160)	0.104 (0.159)	0.243 (0.094)	0.147 (0.124)
RS	<rmsEmin>	1.815 (0.133)	1.775 (0.118)	1.813 (0.108)	1.818 (0.098)

Tabla 7. Resumen de los resultados del muestreo utilizando tres métodos diferentes, REMC, SA y RS, sobre dos superficies de energía distintas. Entre paréntesis se muestra la desviación estándar.

En la Tabla 7 se aprecia que todas las estructuras de menor energía obtenidas con REMC están muy cerca de la diana, con valores de energía mínima muestreada *Emin*, muy cercanos al mínimo global y valores de *rmsEmin* menores de 0.4 Å. Esto significa que el método EPA-REMC es realmente capaz de encontrar estructuras muy cercanas al mínimo global en las superficies de energía modeladas. Para los métodos SA y RS, los valores de energía mínima

muestreados se quedan más lejos del mínimo global y todas las estructuras predichas tienen RMSD´s de entre 1 y 1.8 Å, lo que indica que las superficies de energía diseñadas son "*difíciles*" para estos métodos de optimización y que el método REMC es más potente que éstos para encontrar el mínimo global en estas superficies de energía multidimensionales. Asimismo, las desviaciones estándar de las energías (mostradas entre paréntesis), para REMC son más pequeñas que los correspondientes valores para SA y RS, lo que implica que REMC es también más estable que los otros dos métodos.

DISCUSIÓN

5. Discusión

El trabajo realizado en esta tesis forma parte del objetivo más amplio de mejorar la calidad de los modelos de estructuras de proteínas propuestos por los métodos de predicción estructural basados en el uso de proteínas plantilla. A pesar de que estos métodos se han beneficiado mucho en los últimos años de las iniciativas de genómica estructural (Sanchez et al., 2000) y han experimentado mejoras considerables en muchas de las etapas de las que constan (Kelley et al., 2000; Koh et al., 2003; Marti-Renom et al., 2004; Shi et al., 2001), el refinado del modelo final propuesto sigue siendo una tarea difícil, tanto, que se ha identificado como un área a desarrollar (Kryshtafovych et al., 2005; Moult, 2005; Tress et al., 2005; Valencia, 2005) si el objetivo es obtener modelos suficientemente buenos como para que puedan ser empleados en experimentos que requieran modelos de alta calidad, como reemplazamiento molecular (Giorgetti et al., 2005), predicción de función (Skolnick et al., 2000) o cribado virtual de fármacos (Lengauer et al., 2004). La dificultad para llevar a cabo un buen refinado se debe al delicado balance de fuerzas en el estado nativo de las proteínas (que todavía no es reproducible en toda su extensión mediante los campos de fuerza actuales), y a la necesidad de muestrear un gran número de conformaciones alternativas en la búsqueda del mínimo global de energía. En esta tesis se abordó esta segunda cuestión.

Intentos previos de refinado estructural incluyeron el uso de dinámica molecular (Lee et al., 2001b; Simmerling et al., 2000) y técnicas basadas en potenciales estadísticos (Lu and Skolnick, 2003), así como el uso de múltiples plantillas (Contreras-Moreira et al., 2003) para tratar de mejorar la calidad final abordando el problema desde sus dos frentes, es decir, tratando de mejorar tanto el muestreo (Contreras-Moreira et al., 2003; Lee et al., 2001b; Offman et al., 2006; Simmerling et al., 2000), como los campos de fuerza (Lu and Skolnick, 2003; Misura et al., 2006). Se demostró (Simmerling et al., 2000) que las simulaciones de dinámica molecular que usan una metodología combinada de muestreo localmente enriquecido (*Locally Enhanced Sampling*, LES) y una red de partículas de Ewald (*Particle Mesh Ewald*, PME), con un campo de fuerza de calidad y un tratamiento explícito del solvente podían llegar a mejorar las estructuras, ya que para el caso estudiado por Simmerling et al (la proteína CMTI-1, de 29 residuos y tres puentes disulfuro), se consiguió una mejora de RMSD de 1.2 Å. Sin embargo, al tratar de trasladar la aplicabilidad del método al caso de proteínas más grandes (Lee et al., 2001b) no se obtuvo ninguna mejora estructural. Lu y Skolnick (Lu and Skolnick, 2003) por su parte, mostraron por primera vez que un procedimiento de refinado podía mejorar la calidad de la estructura final usando potenciales estadísticos basados en restricciones locales (*Local Constraint Refinement*, LCR) y restricciones de contactos (*Reduced Contact Refinement*, RCR). Para estos dos métodos, de un total de 67 casos de refinado, observaron una mejora modesta en aproximadamente la mitad de los casos, un deterioro de la estructura en el 10% de los mismos y

en el resto no se apreciaron cambios significativos. Asimismo, observaron que el uso combinado de los potenciales estadísticos con dinámica molecular también podía mejorar modestamente las estructuras, pero sólo en el caso de proteínas α. Aunque los resultados fueron bastante modestos, consiguieron mejoras sobre los estudios previos que sólo usaban dinámica molecular (Lee et al., 2001b) y lograron poner de manifiesto que era posible aplicar de manera consistente un procedimiento de refinado para mejorar la calidad de las estructuras. No obstante, el objetivo de producir modelos de resolución atómica todavía estaba lejos de conseguirse.

Posteriormente, una metología que combinaba la función de energía de Rosetta (Bonneau et al., 2002; Bradley et al., 2003), con un protocolo de refinado "*full atom*" (Misura and Baker, 2005) y restricciones de distancia derivadas de estructuras homólogas, permitió construir modelos que tenían frecuentemente mayor calidad que las plantillas de partida (Misura et al., 2006). Los modelos resultantes eran físicamente realistas, ya que contenían aproximadamente el mismo número de solapamientos atómicos que las estructuras cristalográficas experimentales y mantenían una buena estereoquímica. Asimismo, se podían identificar mediante sus energías y en más de la mitad de los casos, uno de los 10 modelos de menor energía identificados correspondía a un modelo de mayor calidad que la plantilla en las regiones alineadas, demostrando de nuevo que la mejora de la calidad era posible.

Por otro lado, en teoría, usar más de una plantilla debería generar modelos de mayor calidad que los que se obtenían a partir de plantillas individuales, debido a una mayor cobertura del espacio conformacional. Sin embargo, esto sólo se conseguía muy ocasionalmente; en general, los modelos construidos a partir de múltiples plantillas no eran mejores que sus correspondientes modelos basados en la plantilla óptima, es más, podían llegar a ser considerablemente peores (Contreras-Moreira et al., 2003). Lo cual puso de manifiesto que las metodologías no eran capaces de sacar ventaja de la mayor riqueza del espacio conformacional disponible y que era necesario diseñar los movimientos adecuados en este espacio para tratar de obtener mejores estructuras. En este sentido, un protocolo desarrollado recientemente permitió mejorar las conformaciones de los modelos utilizando múltiples plantillas y un conjunto de movimientos ("*move set*"), combinado con un algoritmo genético. Este método permitió muestrear regiones relativamente grandes del espacio conformacional accesibles a la estructura nativa (Offman et al., 2006). Sin embargo, la definición de un conjunto de movimientos óptimo en el espacio, crucial para la efectividad de la técnica, es un problema no resuelto aún.

Por ello, conocer los cambios estructurales que experimentan las proteínas a lo largo de la evolución y utilizar esta información en el problema del refinado, como se pretende en esta tesis, podría ser útil para obtener mejores resultados. La comparación de estructuras conocidas pertenecientes a una familia de proteínas homólogas y la caracterización desde un punto de vista físico de las deformaciones estructurales que han tenido lugar dentro de esa familia, podría permitir la obtención de estructuras de mayor calidad. Ya se demostró que la combinación del

refinado basado en energía, con el muestreo a lo largo de las direcciones observadas en los cambios de las estructuras durante la evolución (PCA), permite superar parcialmente los principales obtáculos que presenta esta etapa (Qian et al., 2004). Qian et al., definieron su espacio de muestreo, como aquél formado únicamente por los tres primeros componentes principales. La utilización de la función de energía de alta resolución de Rosetta (Bonneau et al., 2002; Bradley et al., 2003) permitió identificar modelos de baja energía muestreados en este espacio. Estos modelos presentaban de manera consistente menores valores de RMSD que los patrones de partida utilizados con respecto a la cadena principal de la estructura nativa. Sin embargo, aunque esperanzadoras, las mejoras globales fueron modestas (tan sólo ~0.3 Å de ganancia en promedio). No obstante, no estaba claro hasta qué punto esto se debía a una limitación en el campo de fuerza empleado o al hecho de que sólo se utilizaran 3 componentes principales en los cálculos.

En esta tesis, se estudiaron de manera independiente los efectos de definir el espacio de muestreo, de aquellos derivados del algoritmo de búsqueda y de los de la función de energía empleados. Se desarrolló un nuevo algoritmo de alineamiento estructural múltiple, MAMMOTH-mult (Lupyan et al., 2005) y se usó esta herramienta junto con el análisis de componentes principales y el análisis de modos normales para estudiar las propiedades de plasticidad de las estructuras en familias de proteínas homólogas. Se mostró que las principales deformaciones que tienen lugar en el *centro estructural* de estas proteínas son altamente cooperativas y se dan en un espacio de baja dimensionalidad (4 ó 5 dimensiones) (Leo-Macías et al., 2005). Esto facilita mucho el muestreo ya que estas componentes principales son direcciones muestreadas evolutivamente y representan posibles movimientos concertados de la cadena. Además, al tratarse tan sólo de unas pocas dimensiones, se minimizan los problemas asociados con posibles imprecisiones en la función de energía, debido a que la reducción en el espacio de muestreo elimina posibles falsos atractores.

Los análisis también arrojaron luz acerca de cuáles son los determinantes estructurales del proceso evolutivo en proteínas. La relación observada entre estas deformaciones principales y los modos vibracionales de baja frecuencia accesibles a una topología particular (Leo-Macías et al., 2005), sugirió que los caminos evolutivos de adaptación estructural hacen uso de combinaciones de un número reducido de modos impuestos por la topología. Así, las deformaciones estructurales permitidas a lo largo de la evolución parecen ser sólo aquellas "*poco costosas*" en términos energéticos, es decir, aquellas que suponen un coste de energía suficientemente pequeño (de ahí, su relación con los modos de más baja frecuencia, que son aquellos de menor energía). Por tanto, la propia topología podría ser un factor importante para determinar la historia evolutiva de las proteínas a nivel estructural.

Así pues, estos resultados, sugirieron la construcción de un espacio de muestreo, EPA, usando los autovectores de EM-PCA y ANM, que podría ser utilizado en el refinado de los

modelos obtenidos mediante modelado por homología. Este espacio hace uso de información evolutiva y vibracional, y presenta dos ventajas muy importantes: su baja dimensionalidad y su resolución satisfactoria. Cuando se comparó la calidad en términos de RMSD de los modelos óptimos que se pueden generar en este espacio (es decir, las proyecciones directas de las proteínas diana sobre él), con la de los métodos estándares de modelado por homología (Figura 32B), se observó una mejora de casi el 50% en RMSD con respecto a lo que se venía haciendo hasta ahora. En cuanto al algortimo de búsqueda, se demostró que las optimizaciones de intercambio de réplicas de Monte-Carlo (REMC), en este subespacio son muy eficientes para encontrar soluciones en las proximidades del mínimo global de energía.

Encontrados tanto un espacio (EPA) como un método (REMC) de muestreo óptimos para tratar de resolver el problema del refinado, como parte del trabajo futuro, queda el estudio de la función de energía, algo que no ha podido abordarse aquí. De hecho, después de casi 30 años de esfuerzo, todavía no ha sido posible derivar tal función (Skolnick, 2006), y aunque ha habido algunos avances, como el uso combinado de potenciales empíricos mejorados optimizados mediante el modelado del perfil energético con la selección de estructuras usando algoritmos de agrupamiento (Wang et al., 2004; Zhang and Skolnick, 2004) que han permitido acercar el modelo a la estructura nativa más allá de la plantilla utilizada (Zhang et al., 2005), la obtención del potencial empírico perfecto para predecir estructuras de proteínas sigue siendo una cuestión abierta, ya que entre otras cosas, todavía son necesarias mejoras significativas en el entendimiento y predicción de las interacciones de corto y medio alcance que podrían dictar la topología global (Fang and Shortle, 2005).

Con respecto a las posibles aplicaciones del espacio EPA obtenido en esta tesis, ya se ha demostrado la utilidad de deformar las estructuras de proteínas a lo largo de los modos normales de vibración de más baja frecuencia tanto a la hora de resolver problemas difíciles de reemplazamiento molecular (Suhre and Sanejouand, 2004b), como a la hora de mejorar el ajuste flexible de estructuras de alta resolución en mapas de baja resolución de microscopía electrónica (Falke et al., 2005; Hinsen et al., 2005; Mitra et al., 2005; Tama et al., 2004) y se han desarrollado servidores muy eficaces para ello: elNemo (Suhre and Sanejouand, 2004a) y NORMA (Suhre et al., 2006) (http://www.igs.cnrs-mrs.fr/elnemo y http://www.igs.cnrs-mrs.fr/elnemo/NORMA). El ajuste flexible a los mapas de microscopía electrónica también se ha beneficiado del uso de estructuras deformadas a lo largo de las principales direcciones evolutivas de cambio estructural observadas en familias de proteínas (Velazquez-Muriel et al., 2006). Por tanto, puesto que nuestro espacio EPA es una combinación del espacio evolutivo y el vibracional podría estudiarse hasta qué punto su uso mejoraría los resultados de ambas técnicas. Resultados preliminares en nuestro grupo han sido prometedores en este sentido en su aplicación al reemplazamiento molecular. Asimismo, combinar nuestro espacio EPA con otra técnica experimental, como RMN, podría ser útil a la hora de facilitar la determinación

estructural. Pruebas realizadas usando acoplamientos residuales dipolares (*residual dipolar couplings*, RDC´s) (Tjandra and Bax, 1997) han apuntado resultados esperanzadores en nuestro grupo, al detectar correlaciones subyacentes que podrían usarse eventualmente de manera provechosa.

Finalmente, otra de las posibles líneas futuras de este trabajo podría abarcar el estudio de las relaciones de los cambios estructurales con las mutaciones en la secuencia a lo largo de la evolución. Una vez estudiada la relación entre las deformaciones evolutivas ("*dinámica evolutiva*") y las deformaciones mecánicas ("*dinámica vibracional*"), sería interesante buscar la posible conexión de éstas con las "deformaciones en secuencia" ("*dinámica secuencial*"). En este sentido, se obtuvieron resultados preliminares que revelaron que las proteínas parecen responder a las mutaciones puntuales reacomodando de manera concertada la estructura de los residuos cercanos a la posición de la mutación, en lugar de experimentar cambios estructurales localizados en la posición mutada (datos no publicados).

CONCLUSIONES

6. Conclusiones

Las principales conclusiones que se pueden derivar del trabajo presentado en esta tesis, se pueden resumir en los siguientes puntos:

1. Los movimientos de adaptación estructural experimentados a lo largo de la evolución en las regiones estructuralmente conservadas de familias de proteínas homólogas son altamente cooperativos y tienen lugar en un espacio de baja dimensionalidad. Ello implica que a nivel del esqueleto polipeptídico, en la naturaleza no existe un problema de búsqueda conformacional.

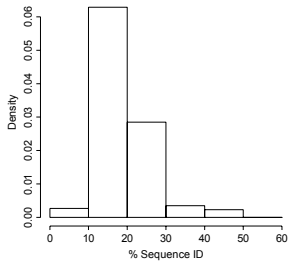
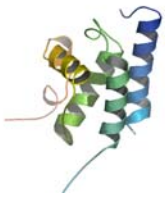
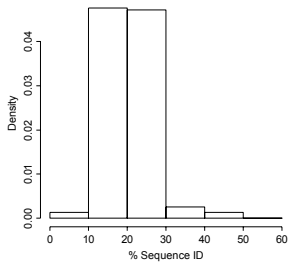
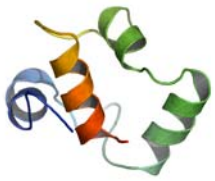
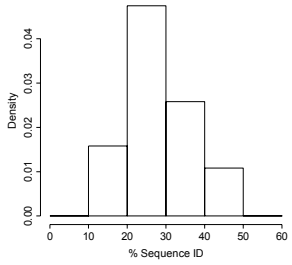
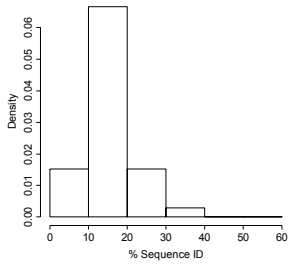
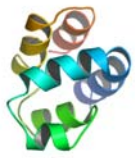
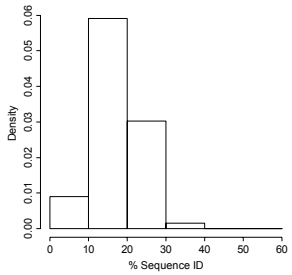
2. Los vectores de deformación mecánica de más baja frecuencia de las proteínas, obtenidos mediante el análisis de sus modos normales de vibración, con un modelo anisotrópico de red elástica, ANM, solapan significativamente con el subespacio evolutivo obtenido mediante EM-PCA. Esto parece indicar que los caminos evolutivos de adaptación estructural de las proteínas hacen uso de los movimientos de baja energía y gran amplitud impuestos por su propia topología.

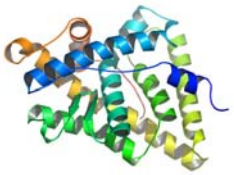
3. Un espacio mixto y de baja dimensionalidad formado por todas las dimensiones de EM-PCA más aquellos vectores de más baja frecuencia de ANM necesarios hasta completar 50 dimensiones, permite representar las estructuras nativas con una buena calidad (por debajo de 1 Å de RMSD en promedio). Este espacio presenta dos ventajas muy importantes: su baja dimensionalidad (que lo hace accesible a las posibilidades computacionales actuales), y una resolución satisfactoria. La baja dimensionalidad permite a su vez, que optimizaciones con el algoritmo de intercambio de réplicas de Monte-Carlo (REMC), puedan encontrar buenas soluciones (ya sea el mínimo global o las proximidades de éste).

MATERIAL SUPLEMENTARIO

7. Material suplementario

Tabla-Mat.Sup. 1. Superfamilias del conjunto de datos 3.1.2 (Ver apartado de Métodos).
a) Todo α..

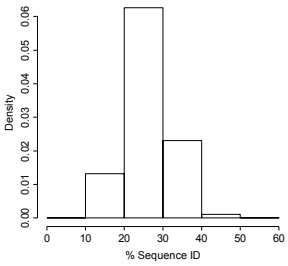
PLEGAMIENTO (Índice de sf.de SCOP)	*Familia SCOP*	*Dominios de ASTRAL*	*% Identidad en sec. (centro)*
 **GLOBINAS (46458)**	**46463**	d3sdha, d1b0b, d1h97a, 1jl7a,d1a6m, d1mba, d1eco, d2gdm,d1irda, d1gcva, d1hjb, d1cg5b,d1gcvb, d1it2a, d1ash, d1itha, d1hlb, d1cqxa1, d1ew6a	
	46459	d1dlwa, d1dlya, d1idra	
	74660	d1kr7a	
 **TRANSFERASAS-S DE LA GLUTACIÓN (47616)**	**47617**	d1glqa1, d2gsta1, d1k3ya1, d1duga1, d1oe8a1, d1ljra1, d1iyha1, d1m0ua1, d2gsq_1, d1eema1, d1e6ba1, d1gwca1, d1oyja1, d1jlva1, d1gnwa1, d1aw9_1, d1a0fa1, d1f2ea1, d1g7oa1, d1k0da1, d1nhya1, d1k0ma1	
 **CITOCROMO C (46626)**	**46680**	d1h1oa1, d1fcdc1, d1fcdc2, detpa2, d1h1oa2	
	46627	d1h32b, d1c53, d1cnoa, d1c52, d451c, d1ql3a, d1ycc, d1i8oa, d1cot	
	68952	d1kb0a1, d1kv9a1	
 **FERRITINA (47240)**	**47241**	d1lkoa1, d1jgca, d1nfva, d1euma, d1jiga, d1o9ra, d1ji4a, d1lb3a	
	47253	d1mtyd, d1mtyb, d1mxra, d1kgna,d1h0oa, d1afra, d1jkva	
 **DOMINIO MUERTE (47986)**	**81312**	d1ngr, d1ddf, d1fada, d1d2za,d1d2zb, d1icha	
	81388	d1a1w, d1n3ka	
	81313	d3crd, d1cy5a, d3ygsp, d1dgna	



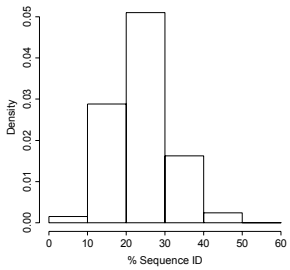
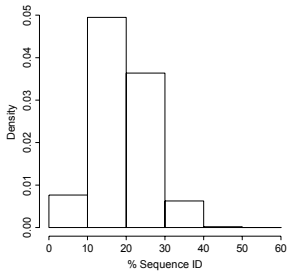
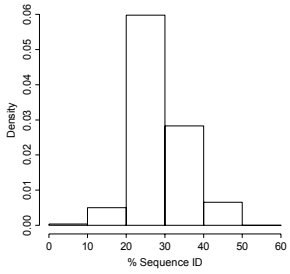
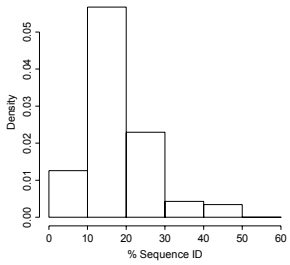
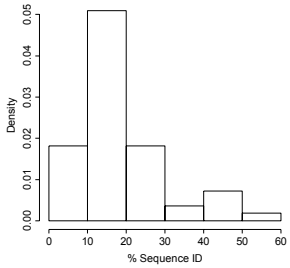
RECEPTOR NUCLEAR DE UNIÓN A LIGANDO (48508)

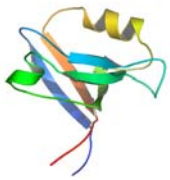
48509

d1fcya, d1pdua, d1a28a, d1qkma, d2prga, d1m13a, d1ie9a, d1n46a, d1g2na, d1n83a, d1kv6a, d1lv2a, d1pk5a, d1p8da



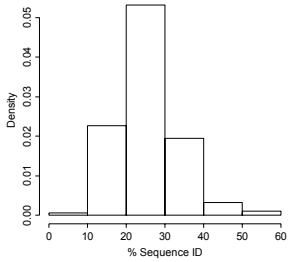
b) Todo β.

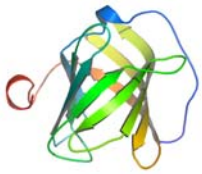
PLEGAMIENTO (Índice de sf.de SCOP)	***Familia SCOP***	***Dominios de ASTRAL***	***% Identidad en sec. (centro)***
 **IMMUNOGLOBULINAS (48726)**	**48942**	d1c5ch2, d1c5cl2, d1dn0b2, d1dr9a2, d1fnga1, d1fngb1, d1fp5a1, d1fp5a2, d1gzqa1, d1hdma1, d1hdmb1, d1hxma2, d1hxmb2, d1hyrc1, d1iam_1,d1k5na1, d1k5nb, d1kgce2, d1l6xa1, d1o0va1, d1vcaa1, d1zxq_1, d2fbjh2	
 **FIBRONECTINA (49265)**	**49266**	d2hft_1, d2hft_2, d1fa, d1fnf_1, d1fnf_2, d1fnf_3, d1fnha1, d1fnha2, d2fnba, d1j8ka, d1qr4a1, d1qr4a2, d1cfb_1, d1cfb_2, d1lwra, d1k85a, d1qg3a1, d1qg3a2, d1axib1, d1axib2, d1eerb1, d1eerb2, d1f6fb1, d1f6fb2, d1iarb1, d1iarb2, d1gh7a1, d1gh7a2, d1gh7a3, d1egja, d1cd9b1, d1cd9b2, d1fyhb1, d1fyhb2, d1bqua1, d1bqua2, d1i1ra1, d1lqsr1, d1lqsr2,d1bpv, d1f42a2, d1f42a3, d1n26a2, d1n26a3, d1n6va1, d1n6va2	
 **SH3 (50044)**	**50045**	d1i07a, d1ng2a1, d1kjwa1, d1pht, d1ckaa, d1awj, d2hsp, d1sema, d1fmk_1, d1gl5a, d1bbza, d1pwt, d1gbra, d1k4us, d1ng2a2, d1oeba, d1bb9, d1i1ja, d1cska, d1neb, d1jqqa, d1ycsb2, d1gcqc, d1jo8a	
 **CUPREDOXINAS (49503)**	**49550** **49504** **63392**	d1kcw_2, d1oe1a2, d1kbva2, d1hfua2, d1aoza2, d1gw0a2, d1kv7a2, d1gska2, d1kv7a3, d1gska3, d1aoza3, d1hfua3, d1gw0a3, d1kcw_5, d1kcw_1, d1kcw_3 d1bqk, d1aac, d1kdj, d1plc, d1bawa d1ibya	
 **GAMMA CRISTALINA (49695)**	**63693** **49710** **49713** **49716** **49719** **49696**	d1g6ea d1wkt d1bhu d1f53a d1c01a d1h4ax1, d1h4ax2, d2bb2_1, d2bb2_2, d1npsa, d1hdfa	



DOMINIO PDZ (50156)

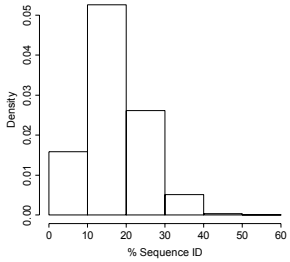
50157	d1kwaa, d1be9a, d1iu0a, d1qava,d1ntea, d1obza1, d1l6oa, d1qaua, d1d5ga, d1g9oa, d1ihja, d1m5za, d1mfga, d1nf3c
68933	d1fc6a3, d1k32a1
74933	d1ky9a1, d1ky9b2, d1lcya1
50172	d1i16

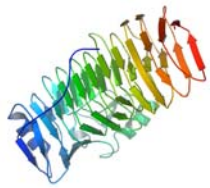




LIPOCALINAS (50814)

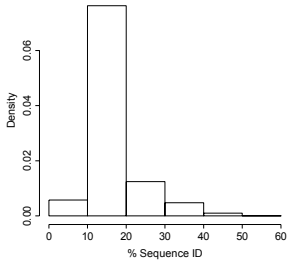
50815	d1kt7a, d1dzka, d1bj7, d1e5pa, d1beba, d1epba, d1jv4a, d1qqsa, d1jzua, d1lf7a, d1kxoa, d1i4ua, d1gkab, d1qfta, d1koia
50847	d1hms, d1ifc, d1mdc, d1crb, d1lfo, d1p6pa, d1o1va
50872	d1avgi





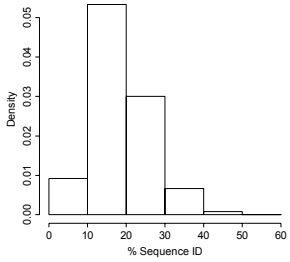
PECTIN-LIASA (51126)

46680	d1h1oa1, d1fcdc1, d1fcdc2, detpa2, d1h1oa2
46627	d1h32b, d1c53, d1cnoa, d1c52, d451c, d1ql3a, d1ycc, d1i8oa, d1co
68952	d1kb0a1, d1kv9a1

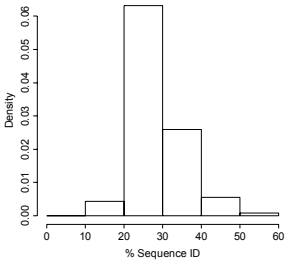
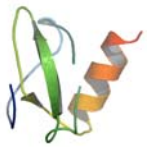
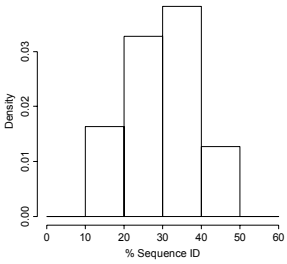
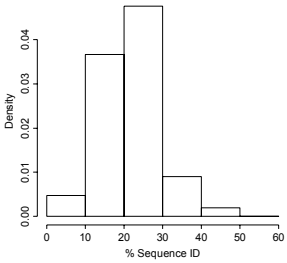
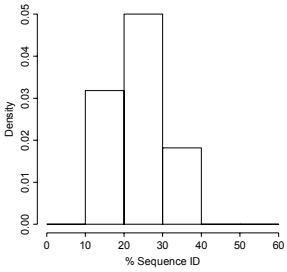
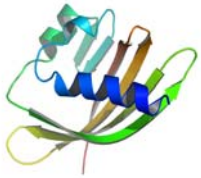
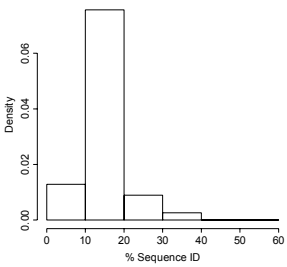


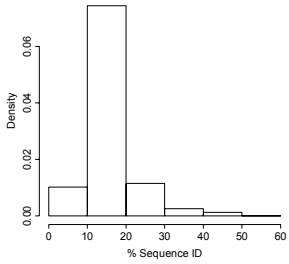
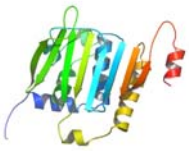
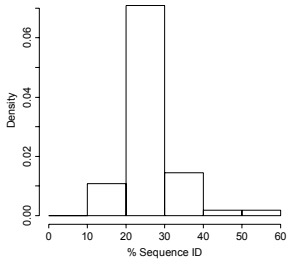
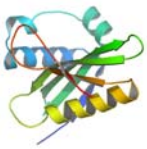
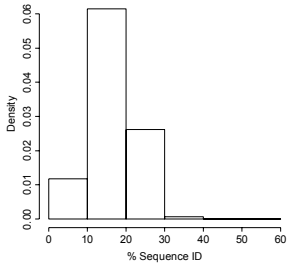
RIBOFLAVINA SINTASA (63380)

51127	d1o88a, d1jtaa, d1bn8a, d1ee6a
51133	d1qcxa
51137	d1rmg, d1bhe, d1k5ca, d1hg8a
51144	d1dbga
69333	d1h80a
51147	d1qjva, d1gq8a
51150	d1qq1a
51153	d1daba

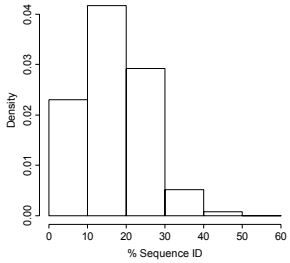
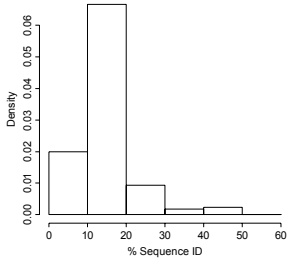
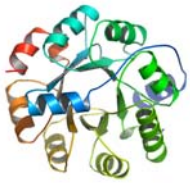
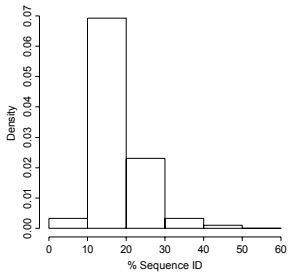
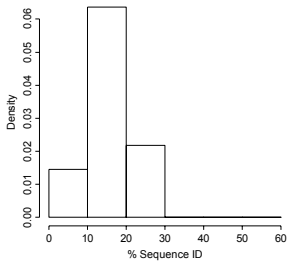
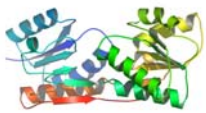
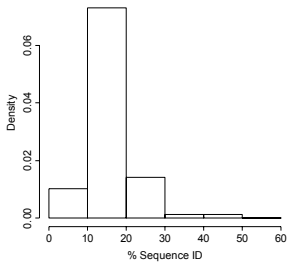


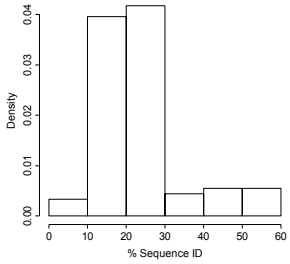
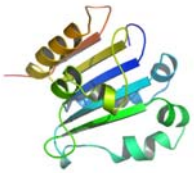
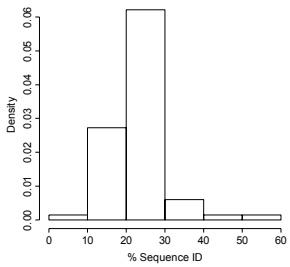
c) α+β

PLEGAMIENTO (Índice de sf.de SCOP)	*Familia SCOP*	*Dominios de ASTRAL*	*% Identitdad en sec.(centro)*
 **QUINASAS (56112)**	**88854**	d1jvpp, d1apme, d1a06, d1kwpa,d1o6la, d1a8a, d1phk, d1gnga,d1kia, d1koba, d1pme, d1csn,d1lpua, d1b6cb, d1f3mc, d1howa,d1jksa, d1o6ya, d1qpca, d1fgka, d1ir3a, d1m14a	
 **QUEMOQUINAS TIPO INTERLEUQUINA 8 (54117)**	**54118**	d1o80a, d1m8aa, d1cm9a, d1b3aa, d1doka, d1el0a, d1eiha, d1g2ta, d1j9oa, d1f2la, d1tvxa	
 **DOMINIO DE UNIÓN A ARN (54928)**	**54929** **54954** **64276**	d1l3ka1, d1l3ka2, d1nu4a, d2u1a, d2u2fa, d1o0pa, d1u2fa, d1fxla1, d1fxla2, d2msta, d1cvja1, d1cvja2, d1qm9a1, d1qm9a2, d1fj7a, d1fjeb2, d1h6kx, d1oo0b, d1owxa d1koha2 d1jmta	
 **DOMINIO C-TERMINAL LDH (56327)**	**56328** **90050**	d7mdha2, d2cmd_2, d1o6za2, d1b8pa2, d1guza2, d1hyha2, d1ldm_2, d1ceqa2, d1ez4a2, d1llda2, d1hyea2 d1obba2	
 **NTF2 (54427)**	**54428** **54431** **89851** **89854** **54434** **54438** **82595** **82598**	d1idpa d1gy7a, d1jkga, d1jkgb, d1of5a,d1of5b d1hkxa d1nwwa d1ocva, d1oh0a d1o7nb d1m98a2 d1m98a2	

 **ABRAZADERA A ADN (55979)**	**55980**	d2pola1, d2pola2, d2pola3	
	55983	d1b77a1, d1b77a2, 1dmla1, d1dmla2, d1plq_1, d1plq_2, d1axca1, d1axca2, d1iz5a1, d1iz5a2	
 **DOMINIO ATPasa DE CHAPERONA HSP90 (55874)**	**55875**	d1byqa	
	55879	d1ei1a2, d1b63a2, d1h7sa2, d1mu5a3	
	55884	d1bxda, d1i58a, d1id0a, d1l0oa	
	69804	d1gkza2, d1jm6a2	
 **ACYL-CoA n-ACILTRANSFERASAS (NAT) (55729)**	**55748**	d1iyka1, d1iyka2	
	55730	d1n71a, d1bo4a, d1m4ia, d1ghea, d1qsta, d1qsma, d1bob, d1fy7a, d1cjwa, d1i12a, d1ufha, d1mk4a, d1nsla	
	75508	d1kzfa	
	82749	d1lrza2, d1lrza3	

d) α / β.

PLEGAMIENTO (Índice de sf.de SCOP)	*Familia SCOP*	*Dominios de ASTRAL*	*% Identidad en sec (centro)*
 **TIOREDOXINA (52833)**	**52855**	d1a8y_1, d1a8y_3, d1a8y_2	
	52849	d1mek, d1bjx, d1a8l_1,	
	52895	d1qgva	
	52892	d1g7ea	
	52834	d1erv, d1fo5a, d1iloa, d1aba, d1qfna, d1kte, d1nm3a1, d1h75a	
	52862	d1k0ma2, d1a0fa2, d1g7oa2, d1ljra2, d1glqa2, d1eema2, d1oyja2, d1jlva2, d1e6ba2, d1gnwa2, d1k0da2, d2gsq_2, d2gsta2, d1k3ya2, d1iyha2, d1nhya2	
 **ALDOLASA (51569)**	**51570**	d1epxa, d1f74a, d1dhpa, d1hl2a, d1euaa, d1n7ka, d1jcla, d1ub3a, d1qfea, d1i2oa, d1l6wa	
	51591	d1dosa, d1gvfa	
	51594	d1l6sa, d1gzga, d1ohla	
	51599	d1jcxa, d1n8fa	
	89494	d1nvma2	
 **BARRIL DE UNIÓN A RIBULOSA-FOSFATO (51366)**	**51367**	d1qo2a, d1thfd	
	51372	d1h1ya	
	51375	d1dbta, d1km3a, d1dqwa, d1kv8a	
	51381	d1pii_2, d1nsj, d1pii_1, d1a53, d1i4na, d1qopa, d1geqa	
 **EXOPEPTIDASAS DEPENDIENTES DE ZN (53187)**	**53188**	d1m4la, d1jqga1, d1h8la2	
	53198	d1obr	
	53201	d1lam_2	
	53204	d1loka, d1qq9a, d1lfwa1, d1cg2a1, d1fnoa4	
	53210	d1de4c3	
 **PROTEÍNA DE UNIÓN PERIPLÁSMICA (53822)**	**53823**	d2dri, d8abp, d1rpja, d1gca, d1pea, d1jx6a, d1dbqa, d1jyea,d1byka, d2liv, d1dp4a, d1jdpa, d1ewka	

 **(FOSFOTYROSINA) FOSFATASAS II (52799)**	**52800**	d1vhra, d1mkp, d1d5ra2, d1ohea1,d1ohea2, d1i9sa, d1fpza	
	52805	d1eeoa, d2shpa1, d1jlna, d1lyva, d1g4us2, d1lara1, d1lara2	
 **FERREDOXINA REDUCTASA, DOMINIO C-TERMINAL DE UNIÓN A NADP (52343)**	**52344**	d1jb9a2, d1fdr_2, d1a8p_2, d1qfja2, d1i7pa2	
	52359	d2pia_2, d1krha2	
	52362	d1ep3b2	
	52365	d1ja1a3, d1f20a2	
	52370	d1cqxa3, d1gvha3	

e) Proteínas pequeñas.

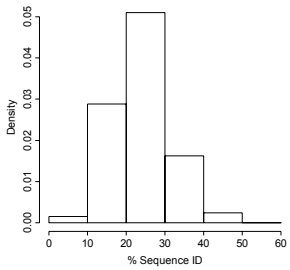
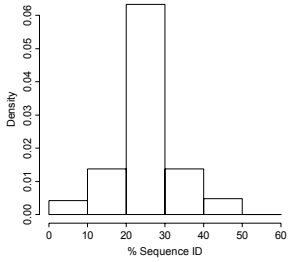
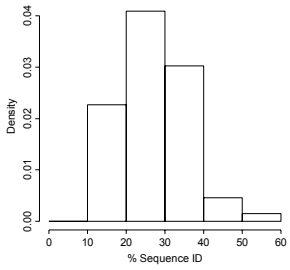
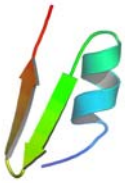
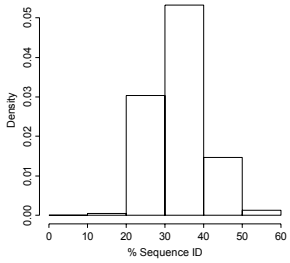
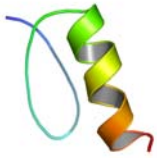
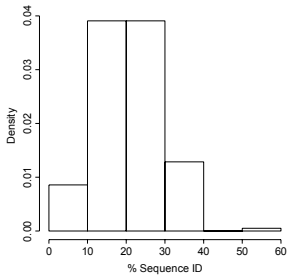
PLEGAMIENTO (Índice de sf.de SCOP)	*Familia SCOP*	*Dominios de ASTRAL*	*% Identidad en sec (centro)*
 **TOXINA DE SERPIENTE (57302)**	**57303**	d2ctx, d1f94a, d3ebx, d1ff4a, d1jgka, d1fas, d1tgxa	
	57354	d1es7b, d1btea, d1m9za, d1erh	
 **DOMINIO SCR (57535)**	**57536**	d1hfi, d1hcc, d1g40a1, d1g40a3, d1g40a4, d1ckla1, d1ckla2, d1quba1, d1quba3, d1quba4, d1quba5, d1h03p1, d1h03p2, 1nwva1, d1gkna1, d1gkna2, d1ly2a1, d1ly2a2, d1elva2, d1gpza2, d1gpza3	
 **DEFENSINA (57392)**	**57393**	d1dfna, d1ijva, d1fd3a, d1kj6a, d1bnb, d1ewsa, d1b8wa, d1d6ba, d2bds, d1sh1, d1atx	
	90157	d1h5oa	
 **TOXINA DE ESCORPIÓN (57095)**	**57096**	d2sn3, d1aho, d1bmr	
	57116	d1lgla, d1jlza, d1scy, d1sis, d1acw, d1sco, d1lir, d1qkya, d1ne5a	
	57160	d1fjna	
	57163	d1i2ua, d1myn, d1ica, d1mm0a	
	57170	d1gps, d1ayj, d1jkza, d1brz, d1jxca	
 **DEDOS DE ZINC C2H2 AND C2HC (57667)**	**57668**	d1a1ia1, d1rmd_1, d2drpa1, d2drpa2, d1paa, d5znf, d1ncs, d2glia1, d2glia2, d2glia4, d1bbo_1, d1bhi, d1ubdc1, d1ubdc2, d1tf3a1, d1tf3a3, d1tf6a6, d1yuja	

Tabla-Mat.Sup. 2. Resumen de los resultados para los modelos *all-atom* construidos a partir de las proyecciones de las dianas en el espacio EPA, antes y después de aplicar una minimización con AMBER8 de los modelos finales. Se detallan los valores del RMSD en Å para los Cα entre los modelos y las estructuras nativas, para toda la estructura modelada, para la región de centro estructural, y para los residuos en lazos. *%Tchi1correct* es el porcentaje de ángulos Chi-1 totales cuya diferencia en valor absoluto entre el modelo y la estructura nativa es menor de 30º. *%Echi1correct* es lo mismo pero sólo para los residuos enterrados (es decir, con una accesibilidad menor del 30%). También se detallan estos resultados para los ángulos Chi-2. En cuanto a los parámetros de Ramachandran, *%RMF* es el porcentaje de residuos en las regiones más favorecidas del mapa; *%RAP* es el porcentaje de residuos en las regiones altamente permitidas; *%RGP* es el porcentaje de residuos en las regiones generosamente permitidas, y *%RD* es el porcentaje de residuos del modelo que están en regiones desfavorecidas del mapa. Por último, *#MalosContc* es el número de malos contactos entre átomos del modelo.

a) Resultados para los modelos minimizados de las "*dianas fáciles"* de CASP5.

Diana	RMSD			Angulos CHI				RAMACHANDRAN				# MALOS CONTC
	Todo	Centro	Lazos	%TCHI-1 correct	%TCHI-2 correct	%ECHI-1 correct	%ECHI-2 correct	%RMF	%RAP	%RGP	%RD	
T0179_2	3,20	1,73	4,70	39,73	42,00	46,43	58,82	61,8	25	9,2	3,9	0
T0138	2,24	1,04	3,80	40,85	46,81	43,33	35,00	70,1	19,5	3,9	6,5	0
T0149_1	3,22	2,54	4,96	27,88	22,67	25,00	22,22	37,9	37,9	17,2	6,9	0
T0150	0,78	0,42	2,82	52,70	50,00	50,00	55,00	87,5	8,8	1,2	2,5	0
T0191_2	2,38	1,64	5,27	44,74	29,03	50,00	38,10	65,1	22,1	11,6	1,2	3
T0169	2,24	1,71	3,75	33,65	27,94	34,09	30,77	56,6	28,3	10,4	4,7	0
T0142	4,01	3,54	5,24	23,76	29,03	18,75	14,29	37,3	37,3	17,6	7,8	3
T0154_1	1,15	0,82	2,57	57,41	44,44	62,50	53,57	70	21,8	3,6	4,5	0
T0136_2	2,31	2,07	4,42	37,40	33,33	38,98	33,33	58,7	23,9	8,7	8,7	2
T0172_1	2,03	1,63	3,75	44,55	27,27	48,89	37,50	63,2	24,5	9,4	2,8	0
T0136_1	2,15	1,79	4,55	44,35	35,53	50,94	31,25	57,6	28	3,8	10,6	1
T0153	1,08	0,38	3,83	52,17	53,57	54,05	52,17	73	23	2	2	0
T0167	2,67	1,33	5,18	42,27	31,88	50,00	35,48	64	19,3	9,6	7	0
T0155	0,98	0,92	2,58	52,38	43,64	41,18	36,84	80,2	13,5	3,1	3,1	0
T0137	1,39	0,71	3,39	46,43	41,67	56,52	48,15	68,8	26,8	2,7	1,8	0
T0183	1,85	1,07	4,45	42,59	38,24	49,02	44,00	63,6	24	5	7,4	0
T0165	2,05	1,80	3,40	25,98	38,20	29,23	55,00	45,1	31,9	12,5	10,4	1
T0185_2	2,02	1,41	4,54	42,11	33,33	46,30	36,36	53,2	27,8	11,9	7,1	3
T0178	3,12	1,24	4,89	43,94	44,09	39,66	50,00	59,2	26,5	10,2	4,1	0
T0189	2,96	2,56	3,82	40,00	27,69	48,89	33,33	31,8	43	18,7	6,5	6
T0182	3,31	1,21	1,46	42,11	41,67	44,78	52,94	74,8	18	4,3	2,9	3

b) Resultados para los modelos minimizados de las "*dianas de dificultad moderada*" de CASP5.

Diana	RMSD			Angulos CHI				RAMACHANDRAN				# MALOS CONTC
	Todo	Centro	Lazos	%TCHI-1 correct	%TCHI-2 correct	%ECHI-1 correct	%ECHI-2 correct	%RMF	%RAP	%RGP	%RD	
T0146_2	2,25	1,09	4,50	41,27	27,66	41,67	33,33	57,7	25,4	11,3	5,6	0
T0162_1	3,71	1,24	3,41	51,52	36,00	50,00	50,00	67,6	14,7	2,9	14,7	0
T0187_2	2,93	2,16	5,56	36,84	35,90	31,58	36,36	35,5	35,5	12,9	16,1	0
T0132	2,53	1,81	4,11	39,62	31,25	36,36	16,67	49,1	34,5	12,7	3,6	0
T0185_3	2,86	0,87	4,72	36,54	35,14	43,75	75,00	54,7	35,8	7,5	1,9	0
T0187_1	1,65	1,37	3,52	36,21	34,09	46,67	60,00	75,4	14,5	5,8	4,3	4
T0184_2	2,74	0,83	5,79	42,31	26,32	57,14	33,33	61,5	25	5,8	7,7	0
T0188	1,82	1,25	4,54	54,72	35,29	61,11	71,43	73,2	16,1	7,1	3,6	1
T0159_1	3,11	1,49	4,22	45,33	30,43	33,33	38,46	44,2	33,8	13	9,1	0
T0130	2,40	2,10	3,48	26,98	29,17	18,18	35,29	40,6	31,2	18,8	9,4	0
T0148_1	2,39	2,24	4,76	37,70	36,36	28,00	40,00	41	31,1	13,1	14,8	0
T0157	2,85	2,33	4,38	45,00	21,88	54,55	14,29	30,2	46,5	16,3	7	1
T0148_2	3,26	1,69	5,72	23,81	40,00	20,00	40,00	33,3	40,9	18,2	7,6	0
T0177_3	2,60	1,82	5,54	27,59	24,00	28,57	26,67	34,5	31	22,4	12,1	8
T0162_3	3,28	2,52	4,01	29,85	46,51	28,57	33,33	52,9	31,4	8,6	7,1	0
T0174_2	2,25	1,84	3,78	35,00	41,89	36,73	45,71	42,5	38,1	8	11,5	1
T0143_2	2,86	1,62	6,14	44,93	40,48	50,00	38,46	54,8	34,2	6,8	4,1	0
T0173	2,15	2,05	3,01	44,16	40,00	45,95	29,41	35,6	41,4	16,1	6,9	0
T0151	2,36	0,59	5,82	47,69	42,11	43,48	50,00	77	16,2	1,4	5,4	0
T0185_1	2,83	1,80	5,88	34,33	40,82	48,28	31,58	60,9	24,6	11,6	2,9	1
T0191_1	4,19	2,64	3,68	26,00	28,21	22,22	25,00	34,5	32,7	21,8	10,9	21
T0154_2	2,35	1,63	4,00	29,51	32,56	40,91	27,78	64,6	20,7	8,5	6,1	0
T0143_1	2,93	2,51	5,23	34,12	36,21	32,35	38,10	51,1	23,9	14,8	10,2	0
T0184_1	3,54	3,58	1,40	40,00	41,66	no apliq	no apliq	78,6	10,7	0	10,7	1
T0193_2	2,55	2,02	4,64	24,00	25,64	35,29	46,15	32,7	41,8	20	5,5	6
T0147	2,66	2,34	2,72	37,66	31,48	34,62	47,37	33,3	41,9	20,4	4,3	2
T0186_2	2,74	2,22	4,76	35,96	32,93	40,00	28,57	41,7	37,5	10,8	10	1

c) Resultados para los modelos minimizados de las "*dianas difíciles*" de CASP5.

Diana	RMSD			Angulos CHI				RAMACHANDRAN				# MALOS CONTC
	Todo	Centro	Lazos	%TCHI-1 correct	%TCHI-2 correct	%ECHI-1 correct	%ECHI-2 correct	%RMF	%RAP	%RGP	%RD	
T0141	2,54	2,62	1,54	32,50	33,33	50,00	50,00	43,2	25	18,2	13,6	0
T0179_1	1,66	1,66	3,42	46,88	40,00	50,00	28,57	46,9	25	12,5	15,6	0
T0172_2	2,76	1,54	6,60	40,38	34,15	27,78	21,43	53,6	28,6	8,9	8,9	2
T0181	3,82	3,47	4,67	19,64	29,73	19,05	42,86	32,7	34,5	23,6	9,1	4
T0162_2	2,58	0,93	3,81	33,33	21,74	33,33	40,00	40,7	18,5	29,6	11,1	0
T0186_1	2,61	1,93	4,77	20,59	30,43	50,00	60,00	24,2	51,5	15,2	9,1	0
T0156	3,57	2,88	5,87	27,91	44,00	38,46	44,44	30,8	42,3	15,4	11,5	4
T0168_1	5,06	3,03	4,97	40,00	32,69	44,00	35,29	39,7	42,3	7,7	10,3	1
T0174_1	3,79	3,43	5,11	27,78	30,91	32,00	31,25	42,9	29,9	11,7	15,6	1
T0193_1	1,30	1,09	2,21	51,43	34,62	60,00	33,33	70,3	16,2	2,7	10,8	0
T0168_2	1,81	0,89	3,83	52,73	52,78	56,25	62,50	73,6	17	3,8	5,7	0
T0161	2,79	2,84	1,90	25,00	33,33	25,00	55,56	36,4	45,5	6,8	11,4	0
T0146_1	2,05	1,40	3,46	33,33	41,18	64,29	25,00	54,7	30,2	7,5	7,5	0
T0177_2	2,44	1,88	4,08	32,50	27,59	28,57	20,00	61	22	12,2	4,9	0
T0177_1	1,94	1,47	5,51	41,38	12,50	40,00	25,00	38,2	29,4	14,7	17,6	2
T0170	1,80	0,25	5,34	45,65	21,62	42,11	22,22	76,5	13,7	5,9	3,9	0
T0176	2,05	0,32	4,63	45,00	21,43	75,00	33,33	66,7	23,8	9,5	0	0
T0159_2	2,18	1,19	4,07	53,42	38,00	59,38	60,00	46,2	37,2	9	7,7	0
T0149_2	3,08	2,16	5,44	25,00	40,00	16,67	42,86	33,3	42,9	11,9	11,9	1

d) Resultados para los modelos sin minimizar de las "*dianas fáciles*" de CASP5.

Diana	RMSD			Angulos CHI				RAMACHANDRAN				# MALOS CONTC
	Todo	Centro	Lazos	%TCHI-1 correct	%TCHI-2 correct	%ECHI-1 correct	%ECHI-2 correct	%RMF	%RAP	%RGP	%RD	
T0179_2	2,93	1,12	4,88	47,95	46,00	53,57	70,59	53,5	19,7	16,9	9,9	42
T0138	2,22	0,95	3,68	42,25	34,04	43,33	40,00	53,2	24,7	14,3	7,8	35
T0149_1	3,11	2,32	5,09	32,69	34,67	30,56	29,63	35,1	18,9	24,3	21,6	61
T0150	0,78	0,39	2,84	59,46	54,00	56,25	55,00	82,5	12,5	2,5	2,5	18
T0191_2	2,23	1,42	5,23	48,68	33,87	56,67	52,38	54,2	25,3	10,8	9,6	27
T0169	2,02	1,40	3,65	38,46	36,76	43,18	38,46	43,3	21,2	23,1	12,5	46
T0142	2,95	1,50	5,27	32,67	22,58	25,00	14,29	44,6	34,9	12	8,4	60
T0154_1	0,96	0,48	2,58	58,33	46,03	60,71	60,71	75	14,8	6,5	3,7	36
T0136_2	2,26	2,02	4,40	35,77	29,49	32,20	18,18	39,3	31,9	20	8,9	95
T0172_1	1,66	1,06	3,74	47,52	41,56	51,11	43,75	63	27	5	5	64
T0136_1	2,14	1,77	4,58	47,83	42,11	56,60	46,88	50	23,8	16,2	10	90
T0153	1,05	0,35	3,76	52,17	44,64	51,35	52,17	68	20	9	3	20
T0167	2,57	1,05	5,19	45,36	34,78	52,27	48,39	60,4	20,7	10,8	8,1	49
T0155	0,98	0,63	2,56	55,95	40,00	52,94	47,37	70,8	14,6	6,2	8,3	19
T0137	1,34	0,63	3,38	50,89	41,67	58,70	51,85	68,2	20	9,1	2,7	32
T0183	1,71	0,76	4,45	50,00	44,12	54,90	52,00	54,8	22,6	12,2	10,4	63
T0165	1,81	1,52	3,29	37,80	40,45	40,00	45,00	42,6	24,3	14	19,1	82
T0185_2	1,98	1,38	4,47	43,86	29,49	46,30	27,27	43,9	26	14,6	15,4	82
T0178	3,02	0,97	4,81	50,76	46,24	44,83	52,94	64	20,9	8,6	6,5	63
T0189	2,62	1,98	3,84	38,18	26,15	46,67	28,57	47,4	13,7	28,4	10,5	70
T0182	3,17	0,69	1,50	45,11	38,10	47,76	50,00	70,6	17,6	6,6	5,1	25

e) Resultados para los modelos sin minimizar de las "*dianas de dificultad moderada*" de CASP5.

Diana	RMSD			Angulos CHI				RAMACHANDRAN				# MALOS CONTC
	Todo	Centro	Lazos	%TCHI-1 correct	%TCHI-2 correct	%ECHI-1 correct	%ECHI-2 correct	%RMF	%RAP	%RGP	%RD	
T0146_2	2,25	1,09	4,50	39,68	38,30	41,67	44,44	40,8	31	18,3	9,9	25
T0162_1	3,55	0,26	3,40	57,58	44,00	66,67	0,00	68,8	15,6	12,5	3,1	1
T0187_2	2,56	1,52	5,55	40,35	33,33	36,84	36,36	20,3	28,8	30,5	20,3	55
T0132	2,17	1,13	4,15	50,94	37,50	54,55	0,00	41,2	23,5	23,5	11,8	35
T0185_3	2,83	0,82	4,84	44,23	27,03	53,33	57,14	64,2	18,9	5,7	11,3	17
T0187_1	1,07	0,40	3,52	36,21	31,82	53,33	70,00	65,2	19,7	6,1	9,1	32
T0184_2	2,65	0,34	5,27	42,31	28,95	57,14	33,33	58,8	19,6	15,7	5,9	17
T0188	1,40	0,40	4,50	60,38	47,06	72,22	57,14	79,2	17	1,9	1,9	21
T0159_1	3,14	1,59	4,22	46,67	39,13	33,33	30,77	28,6	35,1	20,8	15,6	58
T0130	2,37	2,05	3,56	28,57	35,42	18,18	35,29	10,9	31,2	37,5	20,3	58
T0148_1	2,29	2,19	4,54	39,34	31,82	28,00	33,33	24,6	29,5	29,5	16,4	51
T0157	2,06	0,95	4,30	52,50	28,13	63,64	14,29	30,8	33,3	17,9	17,9	26
T0148_2	3,22	1,62	5,74	33,33	44,44	40,00	60,00	25	29,7	25	20,3	41
T0177_3	2,50	1,64	5,51	22,41	18,00	23,81	33,33	24,6	35,1	21,1	19,3	48
T0162_3	2,70	1,49	3,93	34,33	44,19	14,29	66,67	35,3	38,2	17,6	8,8	34
T0174_2	2,22	1,79	3,78	38,00	35,14	36,73	31,43	24,8	22	29,4	23,9	65
T0143_2	2,80	0,86	6,06	42,03	47,62	45,83	23,08	57,7	31	5,6	5,6	31
T0173	2,12	2,01	3,02	48,05	52,50	48,65	52,94	22	28	30,5	19,5	54
T0151	2,34	0,53	5,80	47,69	47,37	43,48	41,67	75	15,3	4,2	5,6	12
T0185_1	2,61	1,28	5,99	35,82	36,73	51,72	36,84	50,7	28,4	16,4	4,5	38
T0191_1	4,04	2,35	3,76	30,00	33,33	22,22	50,00	24,5	37,7	18,9	18,9	46
T0154_2	2,30	1,60	3,94	36,07	34,88	45,45	33,33	37,8	26,8	24,4	11	60
T0143_1	3,06	2,65	5,39	37,65	29,31	35,29	23,81	22,7	26,1	29,5	21,6	74
T0184_1	0,74	0,75	0,41	53,33	37,50	no apliq	no apliq	78,6	14,3	3,6	3,6	5
T0193_2	2,16	1,39	4,64	20,00	17,95	29,41	30,77	40	16	26	18	50
T0147	2,37	1,94	2,74	49,35	44,44	53,85	52,63	25,3	26,4	32,2	16,1	78
T0186_2	2,60	1,98	4,66	42,11	32,93	50,98	31,43	35,1	34,2	15,3	15,3	83

f) Resultados para los modelos sin minimizar de las "*dianas difíciles*" de CASP5.

Diana	RMSD			Angulos CHI				RAMACHANDRAN				# MALOS CONTC
	Todo	Centro	Lazos	%TCHI-1 correct	%TCHI-2 correct	%ECHI-1 correct	%ECHI-2 correct	%RMF	%RAP	%RGP	%RD	
T0141	1,66	1,69	1,61	45,00	40,74	25,00	50,00	38,1	23,8	21,4	16,7	25
T0179_1	1,72	1,70	3,46	46,88	48,00	40,00	42,86	25	31,2	21,9	21,9	26
T0172_2	2,67	1,37	6,55	42,31	43,90	33,33	50,00	43,6	21,8	14,5	20	58
T0181	3,73	3,38	4,71	25,00	16,22	23,81	35,71	27,5	29,4	21,6	21,6	70
T0162_2	2,62	1,07	3,79	36,67	39,13	33,33	60,00	48,1	7,4	18,5	25,9	24
T0186_1	2,57	1,87	4,77	29,41	43,48	50,00	60,00	16,1	19,4	41,9	22,6	19
T0156	3,33	2,55	5,78	37,21	28,00	42,86	55,56	26,5	20,4	34,7	18,4	65
T0168_1	4,72	2,20	4,97	40,00	38,46	44,00	35,29	36,5	28,4	23	12,2	63
T0174_1	3,25	2,69	5,05	29,17	40,00	28,00	31,25	33,3	19,4	22,2	25	45
T0193_1	0,77	0,13	2,09	48,57	30,77	50,00	33,33	72,2	16,7	11,1	0	11
T0168_2	1,61	0,20	3,79	58,18	50,00	56,25	62,50	70	20	6	4	11
T0161	2,52	2,56	1,98	40,00	40,00	41,67	55,56	14,3	33,3	31	21,4	29
T0146_1	1,92	1,11	3,53	26,67	38,24	57,14	25,00	58,8	21,6	7,8	11,8	19
T0177_2	1,70	0,50	4,46	40,00	34,48	42,86	40,00	73	16,2	10,8	0	10
T0177_1	1,62	0,98	5,50	34,48	29,17	40,00	25,00	25	21,9	40,6	12,5	24
T0170	1,82	0,11	5,44	43,48	35,14	47,37	38,89	82,4	5,9	5,9	5,9	31
T0176	2,05	0,18	4,66	45,00	32,14	62,50	33,33	73,8	16,7	9,5	0	17
T0159_2	1,91	0,28	4,03	52,05	44,23	53,13	45,45	53,4	21,9	19,2	5,5	39
T0149_2	2,92	1,93	5,40	38,64	36,67	25,00	42,86	17,5	37,5	37,5	7,5	41

BIBLIOGRAFÍA

8. Bibliografía

Abagyan, R.A., andS. Batalov. 1997. Do aligned sequences share the same fold? *J Mol Biol* 273(1):355-368.

Alberts, B.B., Dennis; Lewis, Julian; Raff, Martin; Roberts, Keith; Watson, James D. . 1994. Molecular Biology of the Cell. Garland Publishing, New York.

Altschul, S.F., andB.W. Erickson. 1985. Significance of nucleotide sequence alignments: a method for random sequence permutation that preserves dinucleotide and codon usage. *Mol Biol Evol* 2(6):526-538.

Altschul, S.F., T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, andD.J. Lipman. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25(17):3389-3402.

Amadei, A., M.A. Ceruso, andA. Di Nola. 1999. On the convergence of the conformational coordinates basis set obtained by the essential dynamics analysis of proteins' molecular dynamics simulations. *Proteins* 36(4):419-424.

Andreeva, A., D. Howorth, S.E. Brenner, T.J. Hubbard, C. Chothia, andA.G. Murzin. 2004. SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Res* 32(Database issue):D226-229.

Anfinsen, C.B. 1973. Principles that govern the folding of protein chains. *Science* 181(96):223-230.

Arfken, G. 1985. Gram-Schmidt Orthogonalization. *In* Mathematical Methods for Physicists. Press A, editor, Orlando. 516-520.

Atilgan, A.R., S.R. Durell, R.L. Jernigan, M.C. Demirel, O. Keskin, andI. Bahar. 2001. Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophys J* 80(1):505-515.

Austin, R.H., K.W. Beeson, L. Eisenstein, H. Frauenfelder, andI.C. Gunsalus. 1975. Dynamics of ligand binding to myoglobin. *Biochemistry* 14(24):5355-5373.

Bahar, I., A.R. Atilgan, andB. Erman. 1997. Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Fold Des* 2(3):173-181.

Bairoch, A., R. Apweiler, C.H. Wu, W.C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M.J. Martin, D.A. Natale, C. O'Donovan, N. Redaschi, andL.S. Yeh. 2005. The Universal Protein Resource (UniProt). *Nucleic Acids Res* 33(Database issue):D154-159.

Baker, D., andA. Sali. 2001. Protein structure prediction and structural genomics. *Science* 294(5540):93-96.

Barton, G.J., andM.J. Sternberg. 1987. A strategy for the rapid multiple alignment of protein sequences. Confidence levels from tertiary structure comparisons. *J Mol Biol* 198(2):327-337.

Bateman, A., L. Coin, R. Durbin, R.D. Finn, V. Hollich, S. Griffiths-Jones, A. Khanna, M. Marshall, S. Moxon, E.L. Sonnhammer, D.J. Studholme, C. Yeats, andS.R. Eddy. 2004. The Pfam protein families database. *Nucleic Acids Res* 32(Database issue):D138-141.

Bates, P.A., L.A. Kelley, R.M. MacCallum, andM.J. Sternberg. 2001. Enhancement of protein modeling by human intervention in applying the automatic programs 3D-JIGSAW and 3D-PSSM. *Proteins* Suppl 5:39-46.

Blundell, T.L., andM.S. Johnson. 1993. Catching a common fold. *Protein Sci* 2(6):877-883.

Blundell, T.L., B.L. Sibanda, M.J. Sternberg, andJ.M. Thornton. 1987. Knowledge-based prediction of protein structures and the design of novel molecules. *Nature* 326(6111):347-352.

Bonneau, R., andD. Baker. 2001. Ab initio protein structure prediction: progress and prospects. *Annu Rev Biophys Biomol Struct* 30:173-189.

Bonneau, R., C.E. Strauss, C.A. Rohl, D. Chivian, P. Bradley, L. Malmstrom, T. Robertson, andD. Baker. 2002. De novo prediction of three-dimensional structures for major protein families. *J Mol Biol* 322(1):65-78.

Bork, P., L. Holm, andC. Sander. 1994. The immunoglobulin fold. Structural classification, sequence patterns and common core. *J Mol Biol* 242(4):309-320.

Bowie, J.U., R. Luthy, andD. Eisenberg. 1991. A method to identify protein sequences that fold into a known three-dimensional structure. *Science* 253(5016):164-170.

Bradley, P., D. Chivian, J. Meiler, K.M. Misura, C.A. Rohl, W.R. Schief, W.J. Wedemeyer, O. Schueler-Furman, P. Murphy, J. Schonbrun, C.E. Strauss, andD. Baker. 2003. Rosetta predictions in CASP5: successes, failures, and prospects for complete automation. *Proteins* 53 Suppl 6:457-468.

Bradley, P., K.M. Misura, andD. Baker. 2005. Toward high-resolution de novo structure prediction for small proteins. *Science* 309(5742):1868-1871.

Brenner, S.E., C. Chothia, andT.J. Hubbard. 1998. Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proc Natl Acad Sci U S A* 95(11):6073-6078.

Brenner, S.E., P. Koehl, andM. Levitt. 2000. The ASTRAL compendium for protein structure and sequence analysis. *Nucleic Acids Res* 28(1):254-256.

Brooks, B., andM. Karplus. 1983. Harmonic dynamics of proteins: normal modes and fluctuations in bovine pancreatic trypsin inhibitor. *Proc Natl Acad Sci U S A* 80(21):6571-6575.

Brooks, B.R., R.E. Bruccoleri, B.D. Olafson, D.J. States, S. Swaminathan, andM. Karplus. 1983. Charmm - a Program for Macromolecular Energy, Minimization, and Dynamics Calculations. *Journal of Computational Chemistry* 4(2):187-217.

Bruccoleri, R.E., andM. Karplus. 1987. Prediction of the folding of short polypeptide segments by uniform conformational sampling. *Biopolymers* 26(1):137-168.

Bruccoleri, R.E.K., M. 1985. Chain closure with bond angle variations. *Macromolecules* 18:2676-2773.

Bujnicki, J.M., A. Elofsson, D. Fischer, andL. Rychlewski. 2001. LiveBench-1: continuous benchmarking of protein structure prediction servers. *Protein Sci* 10(2):352-361.

Burke, D.F., C.M. Deane, andT.L. Blundell. 2000. Browsing the SLoop database of structurally classified loops connecting elements of protein secondary structure. *Bioinformatics* 16(6):513-519.

Bystroff, C., andD. Baker. 1998. Prediction of local structure in proteins using a library of sequence-structure motifs. *J Mol Biol* 281(3):565-577.

Canutescu, A.A., andR.L. Dunbrack. 2003. Cyclic coordinate descent: A robotics algorithm for protein loop closure. *Protein Science* 12(5):963-972.

Canutescu, A.A., A.A. Shelenkov, andR.L. Dunbrack, Jr. 2003. A graph-theory algorithm for rapid protein side-chain prediction. *Protein Sci* 12(9):2001-2014.

CASP. 1994. Critical Assessment of techniques for protein Structure Prediction. PROTEINS: Structure, Function and Genetics, Asilomar, California.

Contreras-Moreira, B., I. Ezkurdia, M.L. Tress, andA. Valencia. 2005. Empirical limits for template-based protein structure prediction: the CASP5 example. *FEBS Lett* 579(5):1203-1207.

Contreras-Moreira, B., P.W. Fitzjohn, andP.A. Bates. 2003. In silico protein recombination: Enhancing template and sequence alignment selection for comparative protein modelling. *Journal of Molecular Biology* 328(3):593-608.

Coutsias, E.S., C.; Jacobson, M.; Dill, K. A. 2004. A kinematic view of loop closure. *J Comput Chem* 25:510-528.

Chandonia, J.M., andS.E. Brenner. 2005. Implications of structural genomics target selection strategies: Pfam5000, whole genome, and random approaches. *Proteins* 58(1):166-179.

Chandonia, J.M., andS.E. Brenner. 2006. The impact of structural genomics: expectations and outcomes. *Science* 311(5759):347-351.

Chandonia, J.M., G. Hon, N.S. Walker, L. Lo Conte, P. Koehl, M. Levitt, andS.E. Brenner. 2004. The ASTRAL Compendium in 2004. *Nucleic Acids Res* 32 Database issue:D189-192.

Chandonia, J.M., N.S. Walker, L. Lo Conte, P. Koehl, M. Levitt, andS.E. Brenner. 2002. ASTRAL compendium enhancements. *Nucleic Acids Res* 30(1):260-263.

Chew, L.P., D. Huttenlocher, K. Kedem, andJ. Kleinberg. 1999. Fast detection of common geometric substructure in proteins. *J Comput Biol* 6(3-4):313-325.

Chothia, C., andA.M. Lesk. 1987. Canonical structures for the hypervariable regions of immunoglobulins. *J Mol Biol* 196(4):901-917.

D.A. Case, T.E.C., III, T. Darden, H. Gohlke, R. Luo, K.M. Merz, Jr., A. Onufriev, C. Simmerling, B. Wang and R. Woods. 2005. The Amber biomolecular simulation programs. . *Computat. Chem* 26:1668-1688.

Dayhoff, M.O., W.C. Barker, andL.T. Hunt. 1983. Establishing homologies in protein sequences. *Methods Enzymol* 91:524-545.

Deane, C.M., andT.L. Blundell. 2001. CODA: a combined algorithm for predicting the structurally variable regions of protein models. *Protein Sci* 10(3):599-612.

Deshpande, N., K.J. Addess, W.F. Bluhm, J.C. Merino-Ott, W. Townsend-Merino, Q. Zhang, C. Knezevich, L. Xie, L. Chen, Z. Feng, R.K. Green, J.L. Flippen-Anderson, J. Westbrook, H.M. Berman, andP.E. Bourne. 2005. The RCSB Protein Data Bank: a redesigned query system and relational database based on the mmCIF schema. *Nucleic Acids Res* 33(Database issue):D233-237.

Eddy, S.R. 1998. Profile hidden Markov models. *Bioinformatics* 14(9):755-763.

Elber, R., andM. Karplus. 1987. Multiple conformational states of proteins: a molecular dynamics analysis of myoglobin. *Science* 235(4786):318-321.

Espadaler, J., N. Fernandez-Fuentes, A. Hermoso, E. Querol, F.X. Aviles, M.J. Sternberg, andB. Oliva. 2004. ArchDB: automated protein loop classification as a tool for structural genomics. *Nucleic Acids Res* 32(Database issue):D185-188.

Eswar, N., Madhusudhan, M.S., Marti-Renom, M.A., Sali, A. 2005. ed. v, editor.

Falke, S., F. Tama, C.L. Brooks, 3rd, E.P. Gogol, andM.T. Fisher. 2005. The 13 angstroms structure of a chaperonin GroEL-protein substrate complex by cryo-electron microscopy. *J Mol Biol* 348(1):219-230.

Fang, Q., andD. Shortle. 2005. A consistent set of statistical potentials for quantifying local side-chain and backbone interactions. *Proteins* 60(1):90-96.

Feig, M.K., J.; Brooks III, C.L. 2001. MMTSB Tool Set. *MMTSB NIH Research Resource, The Scripps Research Institute*.

Fischer, D., A. Elofsson, L. Rychlewski, F. Pazos, A. Valencia, B. Rost, A.R. Ortiz, andR.L. Dunbrack, Jr. 2001. CAFASP2: the second critical assessment of fully automated structure prediction methods. *Proteins* Suppl 5:171-183.

Fiser, A., M. Feig, C.L. Brooks, 3rd, andA. Sali. 2002. Evolution and physics in comparative protein structure modeling. *Acc Chem Res* 35(6):413-421.

Fiser, A., andA. Sali. 2003. ModLoop: automated modeling of loops in protein structures. *Bioinformatics* 19(18):2500-2501.

Fitch, W.M., andT.F. Smith. 1983. Optimal sequence alignments. *Proc Natl Acad Sci U S A* 80(5):1382-1386.

Flory, P.J. 1976. Statistical thermodynamics of random networks. *Proc. R. Soc. Lond. A.* 351:351-380.

Gao, H., J. Sengupta, M. Valle, A. Korostelev, N. Eswar, S.M. Stagg, P. Van Roey, R.K. Agrawal, S.C. Harvey, A. Sali, M.S. Chapman, andJ. Frank. 2003. Study of the structural dynamics of the E coli 70S ribosome using real-space refinement. *Cell* 113(6):789-801.

Giorgetti, A., D. Raimondo, A.E. Miele, andA. Tramontano. 2005. Evaluating the usefulness of protein structure models for molecular replacement. *Bioinformatics* 21(suppl_2):ii72-ii76.

Go, N., T. Noguti, andT. Nishikawa. 1983. Dynamics of a small globular protein in terms of low-frequency vibrational modes. *Proc Natl Acad Sci U S A* 80(12):3696-3700.

Go, N.S., H. J. 1970. Ring closure and local conformational deformations of chain molecules. *Macromolecules* 3:178-187.

Godzik, A. 2003. Fold recognition methods. *Methods Biochem Anal* 44:525-546.

Gribskov, M., A.D. McLachlan, andD. Eisenberg. 1987. Profile analysis: detection of distantly related proteins. *Proc Natl Acad Sci U S A* 84(13):4355-4358.

Gumbel, E. 1958. Statistics of extremes. Press CU, editor, New York.

Hayward, S., A. Kitao, andN. Go. 1995. Harmonicity and anharmonicity in protein dynamics: a normal mode analysis and principal component analysis. *Proteins* 23(2):177-186.

Henikoff, J.G., andS. Henikoff. 1996. Using substitution probabilities to improve position-specific scoring matrices. *Comput Appl Biosci* 12(2):135-143.

Henikoff, S., andJ.G. Henikoff. 1992. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A* 89(22):10915-10919.

Henikoff, S., andJ.G. Henikoff. 1994. Position-based sequence weights. *J Mol Biol* 243(4):574-578.

Hinsen, K., N. Reuter, J. Navaza, D.L. Stokes, andJ.J. Lacapere. 2005. Normal mode-based fitting of atomic structure into electron density maps: application to sarcoplasmic reticulum Ca-ATPase. *Biophys J* 88(2):818-827.

Hobohm, U., M. Scharf, R. Schneider, andC. Sander. 1992. Selection of representative protein data sets. *Protein Sci* 1(3):409-417.

Holm, L., andJ. Park. 2000. DaliLite workbench for protein structure comparison. *Bioinformatics* 16(6):566-567.

Holm, L., andC. Sander. 1994. The FSSP database of structurally aligned protein fold families. *Nucleic Acids Res* 22(17):3600-3609.

Holm, L., andC. Sander. 1996. The FSSP database: fold classification based on structure-structure alignment of proteins. *Nucleic Acids Res* 24(1):206-209.

Holm, L., andC. Sander. 1997. Dali/FSSP classification of three-dimensional protein folds. *Nucleic Acids Res* 25(1):231-234.

Holm, L., andC. Sander. 1998. Touring protein fold space with Dali/FSSP. *Nucleic Acids Res* 26(1):316-319.

Hooft, R.W.W., G. Vriend, C. Sander, andE.E. Abola. 1996. Errors in protein structures. *Nature* 381(6580):272-272.

Hukushima, K., andK. Nemoto. 1996. Exchange Monte Carlo method and application to spin glass simulations. *Journal of the Physical Society of Japan* 65(6):1604-1608.

Hukushima, K., H. Takayama, andK. Nemoto. 1996. Application of an extended ensemble method to spin glasses. *International Journal of Modern Physics C-Physics and Computers* 7(3):337-344.

Jacobson, M.P., D.L. Pincus, C.S. Rapp, T.J. Day, B. Honig, D.E. Shaw, andR.A. Friesner. 2004. A hierarchical approach to all-atom protein loop prediction. *Proteins* 55(2):351-367.

Jaroszewski, L., L. Rychlewski, Z. Li, W. Li, andA. Godzik. 2005. FFAS03: a server for profile--profile sequence alignments. *Nucleic Acids Res* 33(Web Server issue):W284-288.

John, B., andA. Sali. 2003. Comparative protein structure modeling by iterative alignment, model building and model assessment. *Nucleic Acids Res* 31(14):3982-3992.

Johnson, R., andD. Wichern. 1998. Applied Multivariate Statistical Analysis. Prentice Hall, Upper Saddle City, New Jersey.

Jones, T.A., andS. Thirup. 1986. Using known substructures in protein model building and crystallography. *Embo J* 5(4):819-822.

Kalos, M.H.W., P.A. 1986. Monte Carlo methods Volume I: Basics. Wiley Interscience, New York.

Kapp, O.H., L. Moens, J. Vanfleteren, C.N. Trotman, T. Suzuki, andS.N. Vinogradov. 1995. Alignment of 700 globin sequences: extent of amino acid substitution and its correlation with variation in volume. *Protein Sci* 4(10):2179-2190.

Karlin, S., andS.F. Altschul. 1990. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc Natl Acad Sci U S A* 87(6):2264-2268.

Karplus, K., C. Barrett, andR. Hughey. 1998. Hidden Markov models for detecting remote protein homologies. *Bioinformatics* 14(10):846-856.

Karplus, K., R. Karchin, J. Draper, J. Casper, Y. Mandel-Gutfreund, M. Diekhans, andR. Hughey. 2003. Combining local-structure, fold-recognition, and new fold methods for protein structure prediction. *Proteins* 53 Suppl 6:491-496.

Kedem, K., L. Chew, andR. Elber. 1999. Unit-vector RMS (URMS) as a tool to analyze molecular dynamics trajectories. *Proteins* 37(4):554-564.

Kelley, L.A., R.M. MacCallum, andM.J. Sternberg. 2000. Enhanced genome annotation using structural profiles in the program 3D-PSSM. *J Mol Biol* 299(2):499-520.

Kendrew, J.C., G. Bodo, H.M. Dintzis, R.G. Parrish, H. Wyckoff, andD.C. Phillips. 1958. A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature* 181(4610):662-666.

Keskin, O., R.L. Jernigan, andI. Bahar. 2000. Proteins with similar architecture exhibit similar large-scale dynamic behavior. *Biophys J* 78(4):2093-2106.

Kirkpatrick, S.G.J., C. D.; Vecchi, M. P. 1983. *Science* 220:671.

Koh, I.Y., V.A. Eyrich, M.A. Marti-Renom, D. Przybylski, M.S. Madhusudhan, N. Eswar, O. Grana, F. Pazos, A. Valencia, A. Sali, andB. Rost. 2003. EVA: Evaluation of protein structure prediction servers. *Nucleic Acids Res* 31(13):3311-3315.

Kolodny, R.G., L; Levitt, M.; Koehl, P. 2005. Inverse kinematics in biology: The protein loop closure problem. *International Journal of Robotics Research* 24:151-163.

Koonin, E.V., Y.I. Wolf, andG.P. Karev. 2002. The structure of the protein universe and genome evolution. *Nature* 420(6912):218-223.

Krogh, A., M. Brown, I.S. Mian, K. Sjolander, andD. Haussler. 1994. Hidden Markov models in computational biology. Applications to protein modeling. *J Mol Biol* 235(5):1501-1531.

Kryshtafovych, A., C. Venclovas, K. Fidelis, andJ. Moult. 2005. Progress over the first decade of CASP experiments. *Proteins* 61 Suppl 7:225-236.

LaCount, M.W., E. Zhang, Y.P. Chen, K. Han, M.M. Whitton, D.E. Lincoln, S.A. Woodin, andL. Lebioda. 2000. The crystal structure and amino acid sequence of dehaloperoxidase from Amphitrite ornata indicate common ancestry with globins. *J Biol Chem* 275(25):18712-18716.

Lander, E.S., L.M. Linton, B. Birren, C. Nusbaum, M.C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J.P. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, N. Stange-Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter, A. Coulson, R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L. Matthews, S. Mercer, S. Milne, J.C. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R.H. Waterston, R.K. Wilson, L.W. Hillier, J.D. McPherson, M.A. Marra, E.R. Mardis, L.A. Fulton, A.T. Chinwalla, K.H. Pepin, W.R. Gish, S.L. Chissoe, M.C. Wendl, K.D. Delehaunty, T.L. Miner, A. Delehaunty, J.B. Kramer, L.L. Cook, R.S. Fulton, D.L. Johnson, P.J. Minx, S.W. Clifton, T. Hawkins, E. Branscomb, P. Predki, P. Richardson, S. Wenning, T. Slezak, N. Doggett, J.F. Cheng, A. Olsen, S. Lucas, C. Elkin, E. Uberbacher, M. Frazier, R.A. Gibbs, D.M. Muzny, S.E. Scherer, J.B. Bouck, E.J. Sodergren, K.C. Worley, C.M. Rives, J.H. Gorrell, M.L. Metzker, S.L. Naylor, R.S. Kucherlapati, D.L. Nelson, G.M. Weinstock, Y. Sakaki, A. Fujiyama, M. Hattori, T. Yada, A. Toyoda, T. Itoh, C. Kawagoe, H. Watanabe, Y. Totoki, T. Taylor, J. Weissenbach, R. Heilig, W. Saurin, F. Artiguenave, P. Brottier, T. Bruls, E. Pelletier, C. Robert, P. Wincker, D.R. Smith, L. Doucette-Stamm, M. Rubenfield, K. Weinstock, H.M. Lee, J. Dubois, A. Rosenthal, M. Platzer, G. Nyakatura, S. Taudien, A. Rump, H. Yang, J. Yu, J. Wang, G. Huang, J. Gu, L. Hood, L. Rowen, A. Madan, S. Qin, R.W. Davis, N.A. Federspiel, A.P. Abola, M.J. Proctor, R.M. Myers, J. Schmutz, M. Dickson, J. Grimwood, D.R. Cox, M.V. Olson, R. Kaul, C. Raymond, N.

Shimizu, K. Kawasaki, S. Minoshima, G.A. Evans, M. Athanasiou, R. Schultz, B.A. Roe, F. Chen, H. Pan, J. Ramser, H. Lehrach, R. Reinhardt, W.R. McCombie, M. de la Bastide, N. Dedhia, H. Blocker, K. Hornischer, G. Nordsiek, R. Agarwala, L. Aravind, J.A. Bailey, A. Bateman, S. Batzoglou, E. Birney, P. Bork, D.G. Brown, C.B. Burge, L. Cerutti, H.C. Chen, D. Church, M. Clamp, R.R. Copley, T. Doerks, S.R. Eddy, E.E. Eichler, T.S. Furey, J. Galagan, J.G. Gilbert, C. Harmon, Y. Hayashizaki, D. Haussler, H. Hermjakob, K. Hokamp, W. Jang, L.S. Johnson, T.A. Jones, S. Kasif, A. Kaspryzk, S. Kennedy, W.J. Kent, P. Kitts, E.V. Koonin, I. Korf, D. Kulp, D. Lancet, T.M. Lowe, A. McLysaght, T. Mikkelsen, J.V. Moran, N. Mulder, V.J. Pollara, C.P. Ponting, G. Schuler, J. Schultz, G. Slater, A.F. Smit, E. Stupka, J. Szustakowski, D. Thierry-Mieg, J. Thierry-Mieg, L. Wagner, J. Wallis, R. Wheeler, A. Williams, Y.I. Wolf, K.H. Wolfe, S.P. Yang, R.F. Yeh, F. Collins, M.S. Guyer, J. Peterson, A. Felsenfeld, K.A. Wetterstrand, A. Patrinos, M.J. Morgan, P. de Jong, J.J. Catanese, K. Osoegawa, H. Shizuya, S. Choi, andY.J. Chen. 2001. Initial sequencing and analysis of the human genome. *Nature* 409(6822):860-921.

Langley, R. 1970. Practical Statistics. Simply explained. Dover, New York.

Laskowski, R.A., M.W. Macarthur, D.S. Moss, andJ.M. Thornton. 1993. Procheck - a Program to Check the Stereochemical Quality of Protein Structures. *Journal of Applied Crystallography* 26:283-291.

Lee, M.R., D. Baker, andP.A. Kollman. 2001a. 2.1 and 1.8 A average C(alpha) RMSD structure predictions on two small proteins, HP-36 and s15. *J Am Chem Soc* 123(6):1040-1046.

Lee, M.R., J. Tsai, D. Baker, andP.A. Kollman. 2001b. Molecular dynamics in the endgame of protein structure prediction. *J Mol Biol* 313(2):417-430.

Leibowitz, N., R. Nussinov, andH.J. Wolfson. 2001. MUSTA--a general, efficient, automated method for multiple structure alignment and detection of common motifs: application to proteins. *J Comput Biol* 8(2):93-121.

Lengauer, T., C. Lemmen, M. Rarey, andM. Zimmermann. 2004. Novel technologies for virtual screening. *Drug Discovery Today* 9(1):27-34.

Leo-Macías, A., P. Lopez-Romero, D. Lupyan, D. Zerbino, andA.R. Ortiz. 2005. An analysis of core deformations in protein superfamilies. *Biophys J* 88(2):1291-1299.

Lesk, A.M., andC. Chothia. 1980. How different amino acid sequences determine similar protein structures: the structure and evolutionary dynamics of the globins. *J Mol Biol* 136(3):225-270.

Levitt, M. 1992. Accurate modeling of protein conformation by automatic segment matching. *J Mol Biol* 226(2):507-533.

Lindahl, E., andA. Elofsson. 2000. Identification of related proteins on family, superfamily and fold level. *J Mol Biol* 295(3):613-625.

Lo Conte, L., B. Ailey, T.J. Hubbard, S.E. Brenner, A.G. Murzin, andC. Chothia. 2000. SCOP: a structural classification of proteins database. *Nucleic Acids Res* 28(1):257-259.

Lo Conte, L., S.E. Brenner, T.J. Hubbard, C. Chothia, andA.G. Murzin. 2002. SCOP database in 2002: refinements accommodate structural genomics. *Nucleic Acids Res* 30(1):264-267.

Lu, H., andJ. Skolnick. 2003. Application of statistical potentials to protein structure refinement from low resolution Ab initio models. *Biopolymers* 70(4):575-584.

Lupyan, D., A. Leo-Macías, andA.R. Ortiz. 2005. A new progressive-iterative algorithm for multiple structure alignment. *Bioinformatics* 21(15):3255-3263.

Luthy, R., J.U. Bowie, andD. Eisenberg. 1992. Assessment of protein models with three-dimensional profiles. *Nature* 356(6364):83-85.

Manocha, D.C., J. 1994. Efficient inverse kinematics for general 6R manipulators. *IEEE Trans. Robotics Automation* 10:648-657.

Marti-Renom, M.A., M.S. Madhusudhan, andA. Sali. 2004. Alignment of protein sequences by their profiles. *Protein Sci* 13(4):1071-1087.

Marti-Renom, M.A., A.C. Stuart, A. Fiser, R. Sanchez, F. Melo, andA. Sali. 2000. Comparative protein structure modeling of genes and genomes. *Annu Rev Biophys Biomol Struct* 29:291-325.

McGuffin, L.J., andD.T. Jones. 2003. Improvement of the GenTHREADER method for genomic fold recognition. *Bioinformatics* 19(7):874-881.

McLachlan, A.D. 1979. Gene duplications in the structural evolution of chymotrypsin. *J Mol Biol* 128(1):49-79.

Melo, F., andE. Feytmans. 1998. Assessing protein structures with a non-local atomic interaction energy. *J Mol Biol* 277(5):1141-1152.

Melo, F., R. Sanchez, andA. Sali. 2002. Statistical potentials for fold assessment. *Protein Sci* 11(2):430-448.

Metropolis, N.R., A. W.; Rosenbluth, M. N.; Teller, A. H.; Teller, E. 1953. *J. Chem. Phys.* 21:1087.

Misura, K.M., andD. Baker. 2005. Progress and challenges in high-resolution refinement of protein structure models. *Proteins* 59(1):15-29.

Misura, K.M., D. Chivian, C.A. Rohl, D.E. Kim, andD. Baker. 2006. Physically realistic homology models built with ROSETTA can be more accurate than their templates. *Proc Natl Acad Sci U S A* 103(14):5361-5366.

Mitra, K., C. Schaffitzel, T. Shaikh, F. Tama, S. Jenni, C.L. Brooks, 3rd, N. Ban, andJ. Frank. 2005. Structure of the E. coli protein-conducting channel bound to a translating ribosome. *Nature* 438(7066):318-324.

Mitsutake, A., Y. Sugita, andY. Okamoto. 2003. Replica-exchange multicanonical and multicanonical replica-exchange Monte Carlo simulations of peptides. I. Formulation and benchmark test. *Journal of Chemical Physics* 118(14):6664-6675.

Mizuguchi, K., C.M. Deane, T.L. Blundell, M.S. Johnson, andJ.P. Overington. 1998a. JOY: protein sequence-structure representation and analysis. *Bioinformatics* 14(7):617-623.

Mizuguchi, K., C.M. Deane, T.L. Blundell, andJ.P. Overington. 1998b. HOMSTRAD: a database of protein structure alignments for homologous families. *Protein Sci* 7(11):2469-2471.

Moult, J. 2005. A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. *Curr Opin Struct Biol* 15(3):285-289.

Moult, J., andM.N. James. 1986. An algorithm for determining the conformation of polypeptide segments in proteins by systematic search. *Proteins* 1(2):146-163.

Mulder, N.J., R. Apweiler, T.K. Attwood, A. Bairoch, A. Bateman, D. Binns, P. Bradley, P. Bork, P. Bucher, L. Cerutti, R. Copley, E. Courcelle, U. Das, R. Durbin, W. Fleischmann, J. Gough, D. Haft, N. Harte, N. Hulo, D. Kahn, A. Kanapin, M. Krestyaninova, D. Lonsdale, R. Lopez, I. Letunic, M. Madera, J. Maslen, J. McDowall, A. Mitchell, A.N. Nikolskaya, S. Orchard, M. Pagni, C.P. Ponting, E. Quevillon, J. Selengut, C.J. Sigrist, V. Silventoinen, D.J. Studholme, R. Vaughan, andC.H. Wu. 2005. InterPro, progress and status in 2005. *Nucleic Acids Res* 33(Database issue):D201-205.

Murzin, A.G., S.E. Brenner, T. Hubbard, andC. Chothia. 1995. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol* 247(4):536-540.

Nagarajaram, H.A., B.V. Reddy, andT.L. Blundell. 1999. Analysis and prediction of inter-strand packing distances between beta-sheets of globular proteins. *Protein Eng* 12(12):1055-1062.

Needleman, S., andC. Wunsch. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol* 48(3):443-453.

Numerical-Recipies-Software. 1992. NUMERICAL RECIPIES IN FORTRAN 77: THE ART OF SCIENTIFIC COMPUTING. Press CU, editor.

Ochagavia, M.E., andS. Wodak. 2004. Progressive combinatorial algorithm for multiple structural alignments: application to distantly related proteins. *Proteins* 55(2):436-454.

Offman, M.N., P.W. Fitzjohn, andP.A. Bates. 2006. Developing a move-set for protein model refinement. *Bioinformatics* 22(15):1838-1845.

Ortiz, A.R., C.E. Strauss, andO. Olmea. 2002. MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison. *Protein Sci* 11(11):2606-2621.

Pearl, F., A. Todd, I. Sillitoe, M. Dibley, O. Redfern, T. Lewis, C. Bennett, R. Marsden, A. Grant, D. Lee, A. Akpor, M. Maibaum, A. Harrison, T. Dallman, G. Reeves, I. Diboun, S. Addou, S. Lise, C. Johnston, A. Sillero, J. Thornton, andC. Orengo. 2005. The CATH Domain Structure Database and related resources Gene3D and DHS provide comprehensive domain family information for genome analysis. *Nucleic Acids Res* 33(Database issue):D247-251.

Pearson, W.R. 1994. Using the FASTA program to search protein and DNA sequence databases. *Methods Mol Biol* 24:307-331.

Perutz, M.F. 1960. Structure of hemoglobin. *Brookhaven Symp Biol* 13:165-183.

Ptitsyn, O.B., andK.L. Ting. 1999. Non-functional conserved residues in globins and their possible role as a folding nucleus. *J Mol Biol* 291(3):671-682.

Qian, B., A.R. Ortiz, andD. Baker. 2004. Improvement of comparative model accuracy by free-energy optimization along principal components of natural structural variation. *Proc Natl Acad Sci U S A*.

Rohl, C.A., C.E. Strauss, D. Chivian, andD. Baker. 2004. Modeling structurally variable regions in homologous proteins with rosetta. *Proteins* 55(3):656-677.

Rossmann, M.G., andP. Argos. 1976. Exploring structural homology of proteins. *J Mol Biol* 105(1):75-95.

Rost, B. 1999. Twilight zone of protein sequence alignments. *Protein Eng* 12(2):85-94.

Russell, R.B., andG.J. Barton. 1992. Multiple protein sequence alignment from tertiary structure comparison: assignment of global and residue confidence levels. *Proteins* 14(2):309-323.

Sali, A., andT.L. Blundell. 1990. Definition of general topological equivalence in protein structures. A procedure involving comparison of properties and relationships through simulated annealing and dynamic programming. *J Mol Biol* 212(2):403-428.

Sali, A., andT.L. Blundell. 1993. Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol* 234(3):779-815.

Sanchez, R., U. Pieper, F. Melo, N. Eswar, M.A. Marti-Renom, M.S. Madhusudhan, N. Mirkovic, andA. Sali. 2000. Protein structure modeling for structural genomics. *Nat Struct Biol* 7 Suppl:986-990.

Sanchez, R., andA. Sali. 1997. Advances in comparative protein-structure modelling. *Curr Opin Struct Biol* 7(2):206-214.

Saqi, M.A., R.B. Russell, andM.J. Sternberg. 1998. Misleading local sequence alignments: implications for comparative protein modelling. *Protein Eng* 11(8):627-630.

Sauder, J.M., J.W. Arthur, andR.L. Dunbrack, Jr. 2000. Large-scale comparison of protein sequence alignment algorithms with structure alignments. *Proteins* 40(1):6-22.

Sayle, R.A., andE.J. Milner-White. 1995. RASMOL: biomolecular graphics for all. *Trends Biochem Sci* 20(9):374.

Schwede, T., J. Kopp, N. Guex, andM.C. Peitsch. 2003. SWISS-MODEL: An automated protein homology-modeling server. *Nucleic Acids Res* 31(13):3381-3385.

Sharma, G., M. Badescu, A. Dubey, C. Mavroidis, S.M. Tomassone, andM.L. Yarmush. 2005. Kinematics and workspace analysis of protein based nano-actuators. *Journal of Mechanical Design* 127(4):718-727.

Shatsky, M., R. Nussinov, andH.J. Wolfson. 2004. A method for simultaneous alignment of multiple protein structures. *Proteins* 56(1):143-156.

Shi, J., T.L. Blundell, andK. Mizuguchi. 2001. FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. *J Mol Biol* 310(1):243-257.

Siddiqui, A.S., U. Dengler, andG.J. Barton. 2001. 3Dee: a database of protein structural domains. *Bioinformatics* 17(2):200-201.

Siew, N., A. Elofsson, L. Rychlewski, andD. Fischer. 2000. MaxSub: an automated measure for the assessment of protein structure prediction quality. *Bioinformatics* 16(9):776-785.

Simmerling, C., M.R. Lee, A.R. Ortiz, A. Kolinski, J. Skolnick, andP.A. Kollman. 2000. Combining MONSSTER and LES/PME to predict protein structure from amino acid sequence: Application to the small protein CMTI-1. *Journal of the American Chemical Society* 122(35):8392-8402.

Sippl, M.J. 1990. Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. *J Mol Biol* 213(4):859-883.

Sippl, M.J. 1993. Recognition of errors in three-dimensional structures of proteins. *Proteins* 17(4):355-362.

Sippl, M.J. 1995. Knowledge-based potentials for proteins. *Curr Opin Struct Biol* 5(2):229-235.

Skocaj, D.B., H., Leonardis, A. 2002. A robust PCA algorithm for building representations for panoramic images. *In* Computer Vision - ECCV 2002: 7th European Conference on Computer Vision. Proceedings, Part IV. Copenhagen, Denmark. 761-775.

Skolnick, J. 2006. In quest of an empirical potential for protein structure prediction. *Curr Opin Struct Biol* 16(2):166-171.

Skolnick, J., J.S. Fetrow, andA. Kolinski. 2000. Structural genomics and its importance for gene function analysis. *Nature Biotechnology* 18(3):283-287.

Skolnick, J., andD. Kihara. 2001. Defrosting the frozen approximation: PROSPECTOR--a new approach to threading. *Proteins* 42(3):319-331.

Smith, T.F., andM.S. Waterman. 1981. Identification of common molecular subsequences. *J Mol Biol* 147(1):195-197.

Sowdhamini, R., D.F. Burke, J.F. Huang, K. Mizuguchi, H.A. Nagarajaram, N. Srinivasan, R.E. Steward, andT.L. Blundell. 1998. CAMPASS: a database of structurally aligned protein superfamilies. *Structure* 6(9):1087-1094.

Srinivasan, N., andT.L. Blundell. 1993. An evaluation of the performance of an automated procedure for comparative modelling of protein tertiary structure. *Protein Eng* 6(5):501-512.

Steinmetz, A.C., J.P. Renaud, andD. Moras. 2001. Binding of ligands and activation of transcription by nuclear receptors. *Annu Rev Biophys Biomol Struct* 30:329-359.

Stewart, G.W. 1980. The eficcient generation fo random orthogonal matrices with an application to condition estimation. *SIAM J. Numer. Anal.* 17:403-409.

Suhre, K., J. Navaza, andY.H. Sanejouand. 2006. NORMA: a tool for flexible fitting of high-resolution protein structures into low-resolution electron-microscopy-derived density maps. *Acta Crystallogr D Biol Crystallogr* 62(Pt 9):1098-1100.

Suhre, K., andY.H. Sanejouand. 2004a. ElNemo: a normal mode web server for protein movement analysis and the generation of templates for molecular replacement. *Nucleic Acids Res* 32(Web Server issue):W610-614.

Suhre, K., andY.H. Sanejouand. 2004b. On the potential of normal-mode analysis for solving difficult molecular-replacement problems. *Acta Crystallogr D Biol Crystallogr* 60(Pt 4):796-799.

Tama, F., O. Miyashita, andC.L. Brooks, 3rd. 2004. Flexible multi-scale fitting of atomic structures into low-resolution electron density maps with elastic network normal mode analysis. *J Mol Biol* 337(4):985-999.

Tama, F., andY.H. Sanejouand. 2001. Conformational change of proteins arising from normal mode calculations. *Protein Eng* 14(1):1-6.

Thiel, K.A. 2004. Structure-aided drug design's next generation. *Nat Biotechnol* 22(5):513-519.

Thompson, J.D., F. Plewniak, R. Ripp, J.C. Thierry, andO. Poch. 2001. Towards a reliable objective function for multiple sequence alignments. *J Mol Biol* 314(4):937-951.

Tirion, M.M. 1996. Large Amplitude Elastic Motions in Proteins from a Single-Parameter, Atomic Analysis. *Physical Review Letters* 77(9):1905-1908.

Tjandra, N., andA. Bax. 1997. Direct measurement of distances and angles in biomolecules by NMR in a dilute liquid crystalline medium. *Science* 278(5340):1111-1114.

Topham, C.M., N. Srinivasan, C.J. Thorpe, J.P. Overington, andN.A. Kalsheker. 1994. Comparative modelling of major house dust mite allergen Der p I: structure validation using an extended environmental amino acid propensity table. *Protein Eng* 7(7):869-894.

Tramontano, A., andV. Morea. 2003. Assessment of homology-based predictions in CASP5. *Proteins* 53 Suppl 6:352-368.

Tress, M., I. Ezkurdia, O. Grana, G. Lopez, andA. Valencia. 2005. Assessment of predictions submitted for the CASP6 comparative modeling category. *Proteins* 61 Suppl 7:27-45.

Valencia, A. 2005. Protein refinement: a new challenge for CASP in its 10th anniversary. *Bioinformatics* 21(3):277.

van Vlijmen, H.W., andM. Karplus. 1997. PDB-based protein loop prediction: parameters for selection and methods for optimization. *J Mol Biol* 267(4):975-1001.

Velazquez-Muriel, J.A., M. Valle, A. Santamaria-Pang, I.A. Kakadiaris, andJ.M. Carazo. 2006. Flexible fitting in 3D-EM guided by the structural variability of protein superfamilies. *Structure* 14(7):1115-1126.

Venclovas, C., A. Zemla, K. Fidelis, andJ. Moult. 2001. Comparison of performance in successive CASP experiments. *Proteins* Suppl 5:163-170.

Vitkup, D., E. Melamud, J. Moult, andC. Sander. 2001. Completeness in structural genomics. *Nat Struct Biol* 8(6):559-566.

Wang, K., B. Fain, M. Levitt, andR. Samudrala. 2004. Improved protein structure selection using decoy-dependent discriminatory functions. *BMC Struct Biol* 4:8.

Wang, L.T.C., C. C. 1991. A combined optimization method for solving the inverse kinematics problem of mechanical manipulators. *IEEE Trans. Robotics Automation* 7:489-499.

Watson, J.D., andF.H. Crick. 1953. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* 171(4356):737-738.

Wedemeyer, W.J.S., H. A. 1999. Exact analytical loop closure in proteins using polynomial equations. *J. Comp. Chem.* 20:819-844.

Zemla, A. 2003. LGA: A method for finding 3D similarities in protein structures. *Nucleic Acids Res* 31(13):3370-3374.

Zhang, C., andC. DeLisi. 1998. Estimating the number of protein folds. *J Mol Biol* 284(5):1301-1305.

Zhang, Y., A.K. Arakaki, andJ. Skolnick. 2005. TASSER: an automated method for the prediction of protein tertiary structures in CASP6. *Proteins* 61 Suppl 7:91-98.

Zhang, Y., andJ. Skolnick. 2004. SPICKER: a clustering approach to identify near-native protein folds. *J Comput Chem* 25(6):865-871.

Zhou, H., andY. Zhou. 2002. Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. *Protein Sci* 11(11):2714-2726.

Zhou, H., andY. Zhou. 2005. Fold recognition by combining sequence profiles derived from evolution and from depth-dependent structural alignment of fragments. *Proteins* 58(2):321-328.

ARTÍCULOS

9. Artículos

El trabajo realizado en esta tesis dio lugar a las siguientes publicaciones:

(1). Lupyan D, Leo-Macias A, Ortiz AR. A new progressive-iterative algorithm for multiple structure alignment. Bioinformatics. 2005 Aug 1;21(15):3255-63.

(2). Leo-Macias A, Lopez-Romero P, Lupyan D, Zerbino D, Ortiz AR. Core deformations in protein families: a physical perspective. Biophys Chem. 2005 Apr 1;115(2-3):125-8.

(3). Leo-Macias A, Lopez-Romero P, Lupyan D, Zerbino D, Ortiz AR. An analysis of core deformations in protein superfamilies. Biophys J. 2005 Feb;88(2):1291-9.

(4). Ortiz AR, Gomez-Puertas P, Leo-Macias A, Lopez-Romero P,Lopez-Vinas E, Morreale A, Murcia M, Wang K. Computational approaches to model ligand selectivity in drug design. Curr Top Med Chem. 2006;6(1):41-55. Review.

(5). Han, R., Leo-Macías, A.; Zerbino, D.; Contreras-Moreira, B.; Ortiz, A.R. (*en preparación)*. An efficient conformational sampling method for homology modeling.

Una copia de las cuatro primeras se adjunta en las siguientes páginas.